

EXPERIMENTAL PSYCHOLOGY

A Methodological Approach

EXPERIMENTAL

PSYCHOLOGY

A Methodological Approach

SECOND EDITION

F. J. McGUIGAN

*Professor of Psychology
Hollins College*

PRENTICE-HALL OF INDIA PRIVATE LIMITED
New Delhi, 1969

This Indian Reprint—Rs. 12.00

(Original U. S. Edition—Rs. 61.87)

EXPERIMENTAL PSYCHOLOGY, 2nd Ed.

by F. J. McGuigan

PRENTICE-HALL INTERNATIONAL, INC., Englewood Cliffs.

PRENTICE-HALL OF INDIA PRIVATE LIMITED, New Delhi.

PRENTICE-HALL INTERNATIONAL INC., London.

PRENTICE-HALL OF CANADA LTD., Toronto.

PRENTICE-HALL OF JAPAN, INC., Tokyo.

© 1960 by Prentice-Hall, Inc., Englewood Cliffs, N. J., U.S.A.
All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publishers.

The export rights of this book are vested solely in the publisher. ~

This Eastern Economy Edition is the only authorised, complete and unabridged photo-offset reproduction of the latest American edition specially published and priced for sale only in Argentina, Bolivia, Brazil, Burma, Cambodia, Ceylon, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Guyana, Haiti, Honduras, Hong Kong, India, Indonesia, Israel, Jamaica, Laos, Malaysia, Nicaragua, Okinawa, Pakistan, Paraguay, Peru, Philippines, Singapore, South Korea, South Vietnam, Surinam, Thailand, Tobago, Trinidad, Turkey, United Arab Republic, Uruguay and Venezuela.

Reprinted in India by special arrangement with Prentice-Hall, Inc., Englewood Cliffs, N. J., U.S.A.

This book has been published with the assistance of the joint Indian American Textbook Programme.

Printed by G. D. Makhija at the India Offset Press, Delhi and
Published by Prentice-Hall of India Private Limited, New Delhi.

To two charming ladies

CONSTANCE AND JOAN

PREFACE

PREFACE TO FIRST EDITION

Experimental psychology was born with the study of sensory processes; it grew as additional topics, such as perception, reaction time, attention, emotion, learning, and thinking, were added. Accordingly, the traditional course in experimental psychology was a course the content of which was accidentally defined by those lines of investigation followed by early experimenters in those fields. But times change, and so does experimental psychology. The present trend is to define experimental psychology not in terms of specific content areas, but rather as a study of scientific methodology generally, and of the methods of experimentation in particular. There is considerable evidence that this trend is gaining ground rapidly.

This book has been written to meet this trend. His methods no longer confined to but a few areas, the experimental psychologist conducts research in almost the whole of psychology—clinical, industrial, social,

military, and so on. To emphasize this point, we have throughout the book used examples of experiments from many fields, illustrative of many methodological points.

In short, then, the point of departure for this book is the relatively new conception of experimental psychology in terms of methodology, a conception which represents the bringing together of three somewhat distinct aspects of science: experimental methodology, statistics, and philosophy of science. We have attempted to perform a job analysis of experimental psychology, presenting the important techniques that the experimental psychologist uses in his everyday work. Experimental methods are the basis of experimental psychology, of course; the omnipresence of statistical presentations in journals attests the importance of this aspect of experimentation. An understanding of the philosophy of science is important to an understanding of what science is, how the scientific method is used, and particularly of where experimentation fits into the more general framework of scientific methodology. With an understanding of the goals and functions of scientific methodology, the experimental psychologist is prepared to function efficiently, avoiding scientifically unsound procedures and fruitless problems.

Designed as it is to be practical in the sense of presenting information on those techniques actually used by the working experimental psychologist, it is hoped for this book that it will help maximize transference of performance from a course in experimental psychology to the type of behavior manifested by the professional experimental psychologist.

ACKNOWLEDGMENTS

I cannot adequately express the appreciation and indebtedness that I feel to all those persons who have helped me, in a wide variety of ways, with this undertaking. Hollins College has been most generous in furnishing an atmosphere conducive to academic endeavor. My students have furnished both valuable criticism of ideas and exposition, and the reinforcement required for the completion of this project. Among those, however, who offered specific suggestions I find that I am particularly indebted to Drs. Allen Calvin, Victor Denenberg, David Duncan, Paul Meehl, Michael Scriven, Kenneth Spence, Lowell Wine, and Mr. John Berserth. I am also appreciative of the work of Charlotte Fisher and Blanche Buterbaugh for typing a readable manuscript out of a series of near-illegible notes.

I am indebted to the various authors and publishers who so kindly permitted me to draw on their sources (as acknowledged in the text), including Professor Sir Ronald A. Fisher, Cambridge, and to Messrs. Oliver and Boyd Ltd., Edinburgh, for permission to reprint pages Nos. 91-92 from their book *The Design of Experiments*; to reprint Table No. V from their book *Statistical Tables of Biological, Agricultural, and Medical Research*; and to reprint Table No. IV from their book *Statistical Methods for Research Workers*.

PREFACE TO SECOND EDITION

It is heartening to note the substantial number of books published that have taken a methodological approach to experimental psychology since the first edition. The increasing acceptance of the "new experimental psychology" has led to the present revision. The major theme of the book has remained as it was, but I have welcomed the opportunity to revise and expand on a fairly large number of points. Perhaps the most apparent of these is the substitution of raw data from actual experiments for the fictitious data originally used, and of the addition of the new chapter on within-subjects designs. In the former connection I attempted to select studies throughout psychology to illustrate the various designs and methodological points. It was, however, extremely difficult to always obtain the appropriate raw data from others so I (apologetically) had to call heavily on my own past research. In general there has been an upgrading of the presentation, for our students are becoming increasingly sophisticated.

Thoughtful advice and suggestions for this revision were extremely helpful and much appreciated from Drs. Victor Denenberg, Edward Simmel, and Lowell Wine. For the assistance in several ways I again want to thank my students, especially Susan Crandell, and Mrs. Louise Lively. To Hollins College and to the University of California at Santa Barbara goes a special expression of gratitude for providing the excellent conditions under which I worked.

F.J.M.

CONTENTS

ONE	An Overview of Experimentation	1
TWO	The Problem	15
THREE	The Hypothesis	35
FOUR	The Experimental Plan	56
FIVE	Experimental Design: The Case of Two Randomized Groups	<u>95</u>
SIX	Experimental Control	119
SEVEN	The Independent and Dependent Variables	144

EIGHT	Experimental Design: The Case of Two-Matched-Groups	162
NINE	Experimental Design: The Case of More Than Two Randomized Groups	193
TEN	Experimental Design: The Factorial Design	245
ELEVEN	Experimental Design: Within-Subjects Designs	289
TWELVE	The Logical Bases of Experimental Inferences	303
THIRTEEN	The Inductive Schema: An Overview of Some Characteristics of Science	318
FOURTEEN	Generalization, Explanation, and Prediction in Experimentation	327
FIFTEEN	Miscellany	349
	Tables of Squares and Square Roots	367
APPENDIX	Answers to Problems	376
	References	383
	Index	393

EXPERIMENTAL PSYCHOLOGY

A Methodological Approach

AN OVERVIEW OF EXPERIMENTATION

THE NATURE OF SCIENCE

One of the main differences between humans and the lower animals is man's greater ability to engage in abstract thinking. For instance, man is more able to survey a number of diverse items and to abstract certain characteristics that they have in common. In attempting to arrive at a general definition of science we might well proceed in such a manner. That is, we might consider the various sciences as a group and abstract the salient characteristics that distinguish them from other disciplines. Figure 1.1 is a schematic representation of the disciplines man studies, rather crudely categorized into three groups (excluding the formal disciplines, mathematics and logic). Within the inner circle we have represented what are commonly called the sciences. The next circle embraces various disciplines that are not usually thought of as sciences, such as the arts and some of the humanities. Outside that circle are yet other disciplines which, for lack of a better term, are designated as metaphysical disciplines.

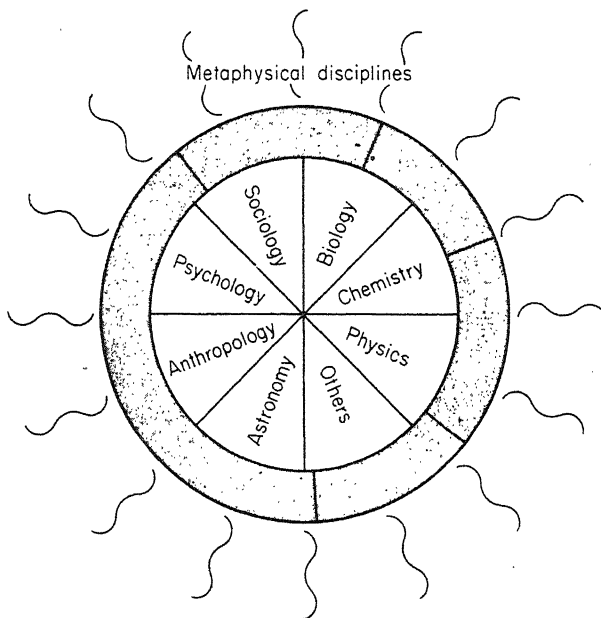


FIGURE 1.1.

Three groups of disciplines which man studies. Within the inner circle are the sciences. The second circle contains the arts and some of the humanities, and the metaphysical disciplines fall outside the circles.

The sciences in the inner circle certainly differ among themselves in a number of ways. But in what important ways are they similar to each other? Likewise, what are the similarities among the disciplines in the outer circle? What do the metaphysical disciplines outside the circle have in common? Furthermore, in what important ways do each of these three groups differ from each other? Answers to these questions should enable us to arrive at an approximation to a general definition of “science.”

One common characteristic of the sciences is that they all use the same general approach in solving problems — “the scientific method.” The “scientific method” is a serial process by which all the sciences obtain answers to their questions. Neither of the other two groups explicitly uses this method.

The disciplines within the two circles differ from the metaphysical disciplines with regard to the type of problem studied. Individuals who study the subject matter areas within the two circles attempt to consider only problems that can be solved; those whose work falls outside of the circle generally study unsolvable problems. Briefly, a *solvable* problem is one that poses a question that can be answered with the use of man’s normal capacities. An *unsolvable* problem raises a question that is essentially unanswerable. Un-

solvable problems usually concern supernatural phenomena or questions about ultimate causes. For example, the problem of what caused the universe is unsolvable and is typical of studies in religion and classical philosophy.¹ Ascertaining what is and what is not a solvable problem is an extremely important topic and will be taken up in greater detail in Chapter Two.

It is important to emphasize that "solvable" and "unsolvable" are technical terms and certain vernacular meanings should not be read into them. It is not meant, for instance, to establish a hierarchy of values among the various disciplines by classifying them according to the type of problem studied. We are not necessarily saying, for example, that the problems of science are "better" or more important than the problems of religion. The distinction is that solvable problems may be attacked by studying observable events in the world around us — they are susceptible to empirical solution, but this is not the case with unsolvable problems. Individuals whose work falls within the two circles (particularly within the inner one) simply believe they must limit their study to problems that they are capable of solving. Of course, some scientists also devote part of their lives to the consideration of supernatural phenomena. But it is important to realize that when they do, they have "left the circle" and are, for that time, no longer behaving as scientists.

In summary, first, the sciences use the scientific method, and they study solvable problems. Second, the disciplines in the outer circle do not use the scientific method, but their problems are typically solvable. And third, the disciplines outside the circles neither use the scientific method nor do they pose solvable problems. These considerations lead to the following definition — "*Science*" is the application of the scientific method to solvable problems. Generally, neither of the other two groups of disciplines have both these features in common.²

With this general definition in hand let us consider the scientific method, primarily as it is applied in psychology. And since the most powerful application of the scientific method is experimentation, we shall focus principally on how experiments are conducted. The problems with which psychologists are concerned are among the most challenging and complex that man faces. For this reason it is necessary to use the most effective methods that science can make available in attempting to solve them. The following brief discussion will serve as an overview of the rest of the book. By obtaining a general picture of how the experimental psychologist proceeds, your orientation to

¹Crude categorizations are dangerous. We merely want to point out *general* differences among the three classes of disciplines. A number of theological problems, for example, are solvable (e.g., "does praying beneficially affect our future behavior?"). Hence, while it is possible to develop at least a limited science of religion, most theologians are not interested in answering their questions in an empirical manner.

²It is likely that there is no completely adequate definition of science available. Although there are limitations to this one, an understanding of it will facilitate presentation of later material.

experimentation should be facilitated. Because this overview is so brief, however, complex matters will necessarily be oversimplified. Possible distortions resulting from this oversimplification will be corrected in later chapters.

**PSYCHOLOGICAL EXPERIMENTATION:
AN APPLICATION OF THE SCIENTIFIC
METHOD³**

A psychological experiment starts with the formulation of a problem, which is usually best stated in the form of a question. The only requirement that the problem must meet is that it be solvable — the question that it raises must be answerable with the tools that are available to the psychologist. Beyond this, the problem may be concerned with any aspect of behavior, whether it is judged to be important or trivial. One lesson of history is that we must not be hasty in judging the importance of the problem on which a scientist works, for many times what was momentarily discarded as being of little importance contributed sizeably to later scientific advances.

The experimenter generally expresses a tentative solution to the problem. This tentative solution is called a hypothesis; it may be a reasoned potential solution or only a vague guess (it is an empirical hypothesis, not a null hypothesis, which will be discussed in Chapter Five). Following the statement of his hypothesis, the experimenter seeks to determine whether the hypothesis is (probably) true or (probably) false, i.e., does it solve the problem he has set for himself? To answer this question he must collect data, for a set of data is his only criterion. Various techniques are available for data collection but, as we said, experimentation is the most powerful.

One of the first steps that the experimenter will take in actually collecting his data is to select a group of subjects with which to work. The type of subject he studies will be determined in part by the nature of the problem. If he is concerned with psychotherapy, he may select a group of mentally disturbed patients. A problem concerned with the function of parts of the brain would entail the use of animals (for few humans volunteer to serve as subjects for brain operations). Learning problems may be investigated with the use of college sophomores, chimpanzees, rats, etc. But whatever the type of subject, the experimenter will assign them to groups. We shall consider here the basic type of experiment, namely, one that involves only two groups.

The assignment of subjects to groups must be made in such a way that the groups will be approximately equivalent at the start of the experiment; this is accomplished through *randomization*, a term to be discussed. The experi-

³There are those who hold that psychologists do not formally go through the following steps of the scientific method in conducting their research. We would agree with this statement for many researchers. However, a close analysis of the actual work of such people would suggest that they at least informally approximate the following pattern, regardless of how they verbalize it.

menter next typically administers an experimental treatment to one of the groups. The experimental treatment is what he wishes to evaluate, and it is administered to the *experimental group*. The other group, called the *control group*, usually receives a normal or standard treatment. It is important, here, to understand clearly just what the terms "experimental" and "normal" or "standard treatment" mean.

In his study of behavior, the psychologist generally seeks to establish empirical relationships between aspects of the environment, broadly conceived, and aspects of behavior. These relationships are known by a variety of names, such as hypotheses, theories, or laws. Such relationships in psychology essentially state that if a certain environmental characteristic is changed, behavior of a certain type also changes.⁴

The aspect of the environment which is experimentally studied is called the *independent variable*; the resulting change in behavior is called the *dependent variable*. Roughly, a variable is anything that can change in value. It is a quality that can exhibit differences in value, usually in magnitude or strength. Thus it may be said that a variable generally is anything that may assume different numerical values. Anything that exists is a variable, according to E. L. Thorndike, for this prominent psychologist asserted that anything that exists, exists in some quantity. Let us briefly elaborate on the concept of a "variable," after which we shall distinguish between independent and dependent variables.

Psychological variables change in value from time to time for any given organism, between organisms, and according to various environmental conditions. Some examples of variables are the height of men, the weight of men, the speed with which a rat runs a maze, the number of trials required to learn a poem, the brightness of a light, the number of words a patient says in a psychotherapeutic interview, and the amount of pay a worker receives for performing a given task.

Figure 1.2 schematically represents one of these examples, "the speed with which a rat runs a maze." It can be seen that this variable can take on any of a large number of magnitudes, or more specifically, it can exhibit any of a large number of time values. In fact, it may "theoretically" assume any of an infinite number of such values, the least being zero seconds, and the

⁴By saying that the psychologist seeks to establish relationships between environmental characteristics and aspects of behavior, we are being unduly narrow. Actually he is also concerned with processes that are not directly observed (variously called logical constructs, intervening variables, hypothetical constructs, etc.). Since, however, it is unlikely that your elementary work will involve hypotheses of such an abstract nature, they will not be further discussed. The highly arbitrary character of defining and differentiating among the various kinds of relationships should be emphasized — frequently the grossly empirical kind of relationship that we are considering under the label "hypothesis," once it is confirmed, is referred to as an empirical or observational law; or before it is tested, merely as a "hunch" or "guess."

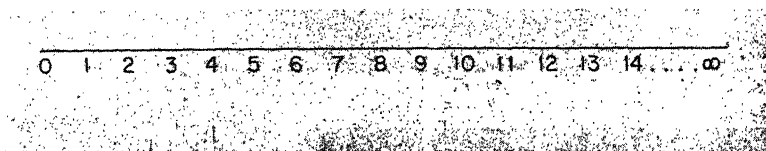


FIGURE 1.2.

Diagrammatic representation of a continuous variable.

greatest being an infinitely large amount of time. In actual situations, however, we would expect it to exhibit a value of a number of seconds, or at the most, several minutes. But the point is that there is no limit to the specific time value that it may assume, for this variable may be expressed in terms of any number of seconds, minutes, hours, including any fraction of these units.

For example, we may find that a rat ran a maze in 24 seconds, in 12.5 seconds, or in 2 minutes and 19.3 seconds. Since this variable may assume any fraction of a value (it may be represented by any point along the line in Figure 1.2), it is called a continuous variable. A continuous variable is one that is capable of changing by any amount, even an infinitesimally small one. A variable that is not continuous is called a discontinuous or discrete variable. A discrete variable can assume only numerical values that differ by clearly defined steps with no intermittent values possible. For example, the number of people in a theatre would be a discrete variable, for, barring an unusually messy affair, one would not expect to find a part of a person in such surroundings. Thus, one might find 1, 15, 299, or 302 people in a theatre, but not 1.6 or 14.8 people. Similarly, gender (male or female), eye color (brown, blue) are frequently cited as examples of discrete variables.⁵

We have said that the psychologist seeks to find relationships between independent and dependent variables. There are an infinite (or at least indefinitely large) number of independent variables available in nature for the psychologist to examine. But he is interested in discovering those relatively few that affect a given kind of behavior. In short, we may say that an independent variable is any variable that is investigated for the purpose of determining whether it influences behavior. Some independent variables that have been scientifically investigated are water temperature, age, hereditary factors, endocrine secretions, brain lesions, drugs, loudness of sounds, and home environments.

Now, with the understanding that an experimenter seeks to determine whether an independent variable affects a dependent variable (either of

⁵We may note that some scientists question whether there are actually any discrete variables in nature. They suggest that we simply "force" nature into "artificial" categories. Color, for example, may more properly be conceived of as a continuous variable — there are many gradations of brown, blue, etc. Nevertheless, scientists find it useful to categorize variables into classes as discrete variables, and to view such categorization as an approximation.

which may be continuous or discrete), let us return to our consideration of experimental and control groups. To determine whether a given independent variable affects behavior the experimenter administers one value of it to his experimental group and a second value of it to his control group. The value administered to the experimental group is, as we have said, the “experimental treatment,” while the control group is usually given a “normal treatment.” Thus, the essential difference between the “experimental” and “normal” treatment is the specific value of the independent variable that is assigned to each group. For example, the independent variable may be the intensity of a shock (a continuous variable). The experimenter may subject the experimental group to a high intensity and the control group to a zero intensity.

To elaborate on the nature of an independent variable, let us consider another example of how one might be used in an experiment. Visualize a continuum similar to Figure 1.2, composed of an infinite number of possible values that the independent variable may take. If, for example, we are interested in determining how well a task is retained as a result of the number of times it is practiced, our continuum would start with zero trials and continue with one, two, three, etc. trials (this would be a discrete variable).

Let us suppose that in a certain industry workers are trained by performing an assembly line task 10 times before being put to work. After awhile, however, it is found that the workers are not assembling their product adequately, and it is judged that they have not learned their task sufficiently well. Some corrective action is indicated, and the foreman suggests that the workers would learn the task better if they were able to practice it 15 times instead of 10. Here we have the makings of an experiment of the simplest sort.

We may think of our independent variable as the “number of times that the task is performed in training,” and will assign it two of the possibly infinite number of values that it may assume — 10 trials and 15 trials. (See Figure 1.3.) Of course, we could have selected any number of other values, one trial, five trials, or 5000 trials, but because of the nature of the problem with which we are concerned, 10 and 15 seem the best values to study. We will

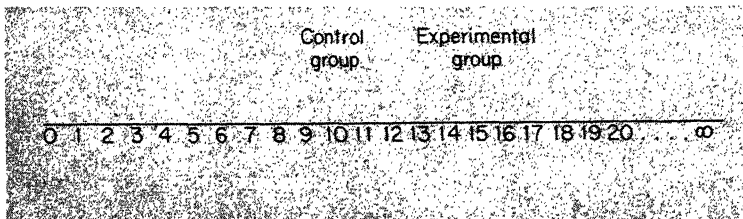


FIGURE 1.3.

Diagrammatic representation of an independent variable. The value of the independent variable assigned to the control group is 10 trials; that assigned to the experimental group, 15 trials.

have the experimental group practice the task 15 times, the control group 10 times. Thus, the control group receives the normal treatment (10 trials), and the experimental group is assigned the experimental or new treatment (15 trials). In many cases, of course, it is arbitrary which group is labeled the control group and which is called the experimental group. Sometimes both treatments are novel ones, in which case it is impossible to label the groups in this manner — they might simply be called “Group 1,” and “Group 2.” In another instance, if one group is administered a “zero” value of the independent variable, and a second group is given some positive amount of that variable, then the zero treatment group would be called the “control group” while the other would be the “experimental group.”

The *dependent* variable is usually some well defined aspect of behavior (a response) which the experimenter measures in his experiment. It may be the number of times the subject says a certain word, the rapidity with which a subject learns a given task, the number of items a worker on a production line can turn out in an hour, and so on. The value obtained for the dependent variable is the criterion of whether or not the independent variable is effective. It is in this sense that it is called a *dependent* variable — the value that it assumes is expected to be dependent on the value assigned to the independent variable.⁶ Thus, an experimenter will vary the independent variable and note whether the dependent variable changes. If the dependent variable changes in value as the independent variable is manipulated, then it may be asserted that there is a relationship between the two. If the dependent variable does not change, however, it may be asserted that there is a lack of relationship between them. For example, let us assume that a light of high intensity is flashed into the eyes of each subject of the experimental group, while those of the control group are subjected to a low intensity light. The dependent variable might be the amount of contraction of the iris diaphragm (the pupil) of the eye which, it may be noted, is an aspect of behavior — a response. If we find that the average contraction of the pupil of the experimental subjects exceeds that of those in the control group, we may conclude that intensity of light is an effective independent variable. We can tentatively assert the following relation: The greater the intensity of a light that is flashed into a subject's eyes, the greater the contraction of the pupil. If, on the other hand, we find no difference in the amount of contraction of the pupil between the two groups, we would assert that there is a lack of relationship between these two variables.

Perhaps the most important principle of experimentation, stated in an ideal form, is that the experimenter must hold constant all of the variables that may affect his dependent variable, except the independent variable(s)

⁶It may be added that the dependent variable is also dependent on some of the *extraneous* variables, discussed shortly, that are always present in an experiment.

that he is attempting to evaluate.⁷ Obviously, there are a number of variables that may affect the dependent variable, but the experimenter is not immediately interested in these. He is, for the moment, interested in only one thing — the relationship, or lack of it, between his independent and his dependent variable. If the experimenter allows a number of other variables to operate in the experimental situation (call them *extraneous* variables) his experiment is going to be contaminated. For this reason he must *control* the extraneous variables in his experiment.

A simple illustration of how an extraneous variable might contaminate an experiment, and thus make the findings unacceptable, might be made using the last example. Suppose that, unknown to the experimenter, all the subjects of his experimental group had that morning received a routine vaccination. But the serum contained a substance that affected the contraction of the pupil of the eye. In this event, the dependent variable measures that the experimenter collected would have, to be generous, little value. For example, if the effect of the serum were such as to cause the pupil not to contract, the subjects of the experimental group would show about the same lack of contraction as the control subjects. It would thus be concluded that the independent variable did not affect the response being studied. The findings would assert that these two variables of light intensity and pupillary contraction are not related, when in fact they are. It was that the dependent variable was affected by an extraneous variable (the serum); and the effects of this extraneous variable obscured the influence of the independent variable. This topic of controlling extraneous variables that might invalidate an experiment is of sufficiently great importance that an entire chapter will be devoted to it. In Chapter Six we shall study various techniques for handling unwanted variables in an experiment.

With this discussion of independent and dependent variables behind us, let us return to our general discussion of the scientific method as applied to experimentation. We have said that a scientist starts his investigation with the statement of a problem, after which he advances a hypothesis as a tentative solution to that problem. He may then conduct an experiment to collect data, data which should indicate the probability that his hypothesis is true or false. He may find it advantageous to use certain types of apparatus and equipment in his experiment. The particular type of apparatus used will naturally depend on the nature of the problem. In general, apparatus is used in an experiment for two main reasons: First, to administer the experimental treatment; and second, to allow, or to facilitate, the collection of data.

The hypothesis that is being tested will predict the way in which the data should point. It may be that the hypothesis will predict that the experi-

⁷Chapter 6 will show why this is an ideal statement and the ways in which it needs to be modified for any given experiment.

mental group will perform better than the control group. By confronting the hypothesis with the dependent variable values of the two groups the experimenter can determine if the hypothesis accurately predicted the results. But it is difficult to tell whether the (dependent variable) values for one group are higher or lower than the values for the second group simply by looking at a number of unorganized data for the two groups of subjects. Therefore, the experimenter must reduce his data to numbers that can be reasonably handled, numbers that will provide him with an answer — for this reason, he must resort to statistics.

For example, he may compute an average (mean) score for both the experimental group and the control group. He might find that the experimental group has a higher mean score (say, 100) than the control group (say, 99). While we note that the experimental group has a higher mean score, we also note that the difference between the two groups is very small. Is this difference, then, a “real” difference, or is it only a chance difference? What are the odds that if we conduct the experiment again, we would obtain the same results? If the difference is a “real,” reliable difference, then the experimental group should obtain a higher mean score than the control group almost every time the experiment is repeated. But if there is no “real” difference between the two groups, we would expect to find the experimental group receiving the higher score half of the time and the control group being superior the other half of the time. To tell whether the difference between the two groups in one experiment is “real,” rather than simply due to random fluctuations (chance), the experimenter resorts to any of a variety of statistical tests. The particular statistical test(s) he uses will be determined by the type of data obtained and the general design of the experiment. But the point is that, on the basis of such tests, it can be determined whether the difference between the two groups is likely to be “real” and reliable or merely “accidental.” More appropriately, the tests indicate whether or not the difference is statistically *significant*, for this is what is meant by a “real” and “reliable” difference. If the difference between the dependent variable scores of the groups is significant, the difference is very probably not due to random fluctuations; and it is concluded that the independent variable is effective (providing that the extraneous variables have been properly controlled).

Thus, by starting with two equivalent groups, administering the experimental treatment to one but not to the other, and collecting and analyzing the (dependent variable) data thus obtained, suppose we find a significant difference between the two groups. We may legitimately assume that the two groups eventually differed because of the experimental treatment. Since this is the result that was predicted by our hypothesis, the hypothesis is supported, or confirmed. In other words, when a hypothesis is supported by experimental data, the probability that the hypothesis is true is increased.

On the other hand, if, in the above example, the control group is found to be equal or superior to the experimental group, the hypothesis is not supported by the data and we may conclude that it is probably false. Naturally, this step of the scientific method in which the hypothesis is tested is considerably oversimplified in our brief presentation. It will be necessary to consider the matter more thoroughly later (See Chapter Twelve).

Closely allied with testing of the hypothesis is an additional step of the scientific method, "generalization." After completing the phases outlined above, the experimenter may feel quite confident that the hypothesis is true for the specific conditions under which he tested it. He must underline *specific* conditions, however, and not lose sight of how specific they are in any given experiment. But the scientist *qua* scientist is not concerned with truth under a highly restricted set of conditions. Rather, he usually wants to make as *general* a statement as he possibly can about nature. And herein lies much of his joy and grief, for the more he generalizes his findings, the greater are the chances for error. Suppose that he has used college students as the subjects for his experiment. This selection does not mean that he is interested *only* in the behavior of college students. Rather, he is probably interested in the behavior of *all* human beings and perhaps even of *all* organisms. Because he has found his hypothesis to be probably true for his particular group of subjects, can he now say that it is probably true for all humans? Or must he simply restrict his results to college students? Or must he narrow the focus even further, limiting it to those students attending the college at which he conducted his experiment? This, essentially, is the question of generalization — how widely can the experimenter generalize his results? He wants to generalize as widely as possible, yet not so widely that the hypothesis "breaks down." The question of how widely he may safely generalize his hypothesis will be discussed in Chapter 14. The broad principle to remember is that he should state that his hypothesis is applicable to as wide a set of conditions (e.g., to as many classes of subjects) as the nature of his experiment warrants.

The next step in the scientific method, closely related to the preceding steps, concerns making predictions on the basis of the hypothesis. By this we mean that a hypothesis may be used to predict certain events in new situations — to predict, for example, that a different group of subjects will act in the same way as a group studied in an earlier experiment.⁸ We can add a final step in the scientific method, *replication*. By *replication* we mean that an additional experiment is conducted in which the *method* of the first experiment is precisely repeated. The confirmed hypothesis may thus be used as the basis for predicting that a new sample of subjects will behave as did the original

⁸A case can be made for not including prediction as a part of the scientific method, at least in some sciences (cf. Scriven, 1959).

sample. If the prediction made by the use of the previously confirmed hypothesis is found to hold in the new situation, the probability that the hypothesis is true is tremendously increased.

In summary let us set down the various steps in the scientific method. (Be advised, however, that there are no rigid rules to follow in doing this. In any process that one seeks to classify into a number of arbitrary categories, some distortion is inevitable. Another source might offer a different classification, while still another one might refuse, quite legitimately, to attempt such an endeavor.)

First, the scientist states a problem that he wishes to investigate. Next, he formulates the hypothesis, a tentative solution to the problem. Third, he collects data relevant to the hypothesis. Following this, he tests the hypothesis by confronting it with the data and makes the appropriate inferences — he organizes the data through statistical means and determines whether the data support or refute the hypothesis. Fifth, assuming that the hypothesis is supported, he may wish to generalize to all things with which the hypothesis is legitimately concerned, in which case he should explicitly state the generality with which he wishes to advance his hypothesis. Sixth, he may wish to make a prediction to new situations, to events not studied in the original experiment. And finally, he may wish to test the hypothesis anew in the novel situation; that is, he might replicate (conduct the experiment with a new sample of subjects) to determine whether the estimate of the probability of his hypothesis can legitimately be increased.

AN EXAMPLE OF A PSYCHOLOGICAL EXPERIMENT

To make the preceding discussion more concrete, consider an example of how an experiment might be conducted from its inception to its conclusion. This example is taken from the area of clinical psychology which, like any applied area, is a methodologically difficult one in which to conduct sound research. Our purpose, however, is primarily to illustrate the application of the preceding principles. Let us assume that a psychotherapist has some serious questions about how best to proceed with his clients in order to effect a "cure" as efficiently as possible. It usually happens that psychotherapists become aware of the "basic" or "real" problems of their clients before the clients themselves do. Thus, they are generally in a position to offer advice to the clients. In our example, however, the psychotherapist is not sure whether offering direct advice is a good procedure to follow. Not only may the client ignore the advice, but he may even react violently against it, thus retarding the therapeutic process. The problem may be stated as follows: Should a psychotherapist authoritatively advise his clients what their problems are and what they should do about them, or should he sit back and let the clients

arrive at their own assessments of the problems and determine for themselves the best action to take? Assume that the psychotherapist believes the latter to be the better procedure. The reasons for or against his opinion need not detain us here. We simply note his hypothesis: If a client undergoing psychotherapy is allowed to arrive at the determination of his problem and its proposed solution by himself, then his recovery will be more efficient than if the psychotherapist gives him this information in an authoritative manner. We might identify the independent variable as "the amount of guidance furnished to the clients," and assign two values to it: first, a maximal amount of guidance; and second, a zero (or at least minimal) amount of guidance.

Suppose that the psychotherapist has ten clients, and that he randomly assigns them to two groups of five each. A great deal of guidance will then be given to one of the groups, and a minimum amount will be administered to the second group. The group that receives a minimum amount of guidance will be called the control group; the group that receives the maximum amount of guidance will be called the experimental group.⁹

Throughout the course of therapy, then, the psychotherapist administers the two different treatments to the experimental and control groups. During this time he prevents the important extraneous variables from acting differently on the two groups. For example, he would want the clients from both groups to undergo therapy in the same place (his office, for instance) to eliminate the possibility that the progress of one group might differ from that of the other group because of the immediate surroundings in which the therapy takes place.

The dependent variable here may be specified as the progress toward recovery. Such a variable is obviously rather difficult to measure, but for illustrative purposes we might use a time measure. Thus, we might assume that the earlier the client is discharged by the therapist, the greater is his progress toward recovery. The time of discharge might be determined when the client has no further complaints of difficulties. Assuming that the extraneous variables have been adequately controlled, the progress toward recovery (the dependent variable) depends on the particular values of the independent variable used, and on nothing else.

As therapy progresses the psychotherapist collects his data. Specifically, he determines the amount of time each client spends in therapy before he is discharged. After all of the clients are discharged, the therapist compares the times for the experimental group against those for the control group. Let us assume that he finds that the mean amount of time in therapy of the experimental group is higher than that of the control group, and further, that a statistical test indicates that the difference is significant. That is, the group

⁹This example well illustrates that *frequently* it is not appropriate to say that a zero amount of the independent variable can be administered to a control group.

that received minimum guidance had a significantly lower time-in-therapy (the dependent variable) than did the group that received maximal guidance. This is precisely what the therapist's hypothesis predicted. Since the results of the experiment are in accord with the hypothesis, we may conclude that the hypothesis is confirmed.

Now the psychotherapist is happy, since he has solved his problem and now knows which of the two methods of psychotherapy is better. But has he found "truth" only for himself, or is what he has found applicable to other situations — can other therapists also benefit by his results? Can his findings be extended, or generalized, to all therapeutic situations of the nature that he has studied? After serious consideration of these matters, he formulates his answer and publishes his findings in a psychological journal.

Inherent in the process of generalization is that of prediction (although there can be generalizations that are not used to make predictions). Here, in effect, what the therapist does, if he generalizes, is to predict that the same results would be obtained if the experiment were repeated in a new situation. In this simple case the therapist would essentially say that for other subjects, offering minimal guidance will result in more rapid recovery than if maximal guidance is offered them. To test this prediction, another psychotherapist might conduct an experiment as outlined above (the experiment is replicated). If his findings prove to be the same, the hypothesis is again supported by the data. With this independent confirmation of the hypothesis as an added factor, it may be concluded that the probability of the hypothesis is increased. That is, our confidence that the hypothesis is true is considerably greater than before.¹⁰

With this overview before us, let us now turn to a detailed consideration of the phases of the scientific method as it applies to psychology. The first matter on which we should enlarge is "the problem."

¹⁰The oversimplification of several topics in this chapter is especially apparent in this fictitious experiment. For one, adequate control would have to be exercised over the important extraneous variable of the therapist's own confidence in, and preference for, one method of psychotherapy. For another, it would have to be demonstrated that the subjects used in this study are typical of those elsewhere before a legitimate generalization of the findings could be asserted. But these problems will be handled in due time. After you finish this book, should you care to concentrate on methodological problems in this area you might read the valuable compendium by Kiesler (1966) entitled "Some Myths of Psychotherapy Research and the Search for a Paradigm."

THE PROBLEM

WHAT IS A PROBLEM?

A scientific inquiry starts when we have already collected a certain amount of knowledge, but all that we can tell from that knowledge is that there is something we don't know. It may be that we simply do not have enough information to answer a question, or it may be that the information that we have is in such a state of disorder that it cannot be adequately related to the question. In either case a problem exists. Let us now see, in a more specific way, how we become aware of a problem.

WAYS IN WHICH A PROBLEM IS MANIFESTED

The lack of sufficient knowledge that bears on a problem is manifested in at least three, to some extent overlapping, ways: first, when there is a noticeable gap in the results of investigations; second, when the results of

several inquiries disagree; and third, when a "fact" exists in the form of a bit of unexplained information. Let us consider each of these in greater detail.

A GAP IN OUR KNOWLEDGE

Probably the most apparent way in which a problem is manifested is when there is a straightforward absence of information; we are aware of what we know and there is simply something that we do not know. If a community group plans to establish a clinic to provide psychotherapeutic services, two natural questions for them to ask are, "What kind of psychotherapy should we offer?" and "Of the different systems of psychotherapy, which is the most effective?" Now these questions are extremely important, but there are few scientifically acceptable studies that provide answers. Here is an apparent gap in our knowledge. Collection of data with a view toward filling this gap is thus indicated.

Students most often conduct experiments in their classes to solve problems of this type. They become curious about why a given kind of behavior occurs, about whether a certain kind of behavior can be produced by a given stimulus, about whether one kind of behavior is related to another kind of behavior, and so forth. Frequently, some casual observation serves as the basis for their curiosity and leads to the formulation of this kind of problem. For example, one student had developed the habit of lowering her head below her knees when she came to a taxing question on an examination. She thought that this kind of behavior facilitated her problem solving ability, and her reasoning was that she thereby "got more blood into her brain." Queer as such behavior might strike you, or queer as it struck her professors (who, as a result, developed a problem of their own; namely, where in the world did she hide the crib notes that she was so obviously studying), such a phenomenon is possible. And there were apparently no relevant data available. Consequently, the students in the class conducted a rather straightforward, if somewhat unusual, experiment: They asked subjects to solve auditorily presented problems as their bodily positions were systematically maneuvered through space.

Similar problems that have been developed by students are: What is the effect of consuming a slight amount of alcohol on motor performance and on problem solving ability? Can the color of the clothes worn by a roommate be controlled through the subtle administration of verbal reinforcements? Do students who major in psychology evidence a higher amount of situational anxiety than those whose major is a "less dynamic" subject? Such problems as these are rather typically selected for the early experiments in a course in Experimental Psychology, and they are quite valuable, at least in helping the

student to learn appropriate methodology. As students read about previous experiments that have been conducted in areas related to the problem that they have chosen, their storehouse of scientific knowledge grows, and their problems become more sophisticated. One cannot help being impressed by the high quality of research conducted by undergraduate students toward the completion of their course in experimental methodology. Fired by their enthusiasm for conducting their own original research, it is not infrequent for them to attempt to solve problems made manifest by contradictory results or by the existence of phenomena for which there is no satisfactory explanation.

CONTRADICTORY RESULTS

To understand how the results of different attempts to solve the same problem may differ, consider three separate experiments that have been reported in psychological journals. All three experiments were very similar, and they all addressed themselves to the following question: "When a person is learning a task, are rest pauses more beneficial if concentrated during the first part of the total practice session or if concentrated during the last part?" For instance, if a person is to spend ten trials in practicing a given task, would his learning be more efficient if his rest pauses were concentrated between the first five trials (early in learning) or between the last five trials (late in learning)? The general design of all three experiments was as follows: One group of subjects practiced a task with rest pauses concentrated during the early part of the practice session. As these subjects continued to practice the task on additional trials, the length of the rest pauses between trials progressively decreased. A second group of subjects practiced the task with progressively increasing rest pauses between trials; as the number of trials on which they practiced the task increased, the amount of rest between trials became larger.

The first experiment indicated that progressively increasing rest periods are superior (Doré and Hilgard, 1938); the second experiment showed that progressively decreasing rest periods led to superior learning (Renshaw and Schwarzbeck, 1938); while the third experiment indicated that the effects of progressively increasing and progressively decreasing rest periods are about the same (Cook and Hilgard, 1949). Why do these three studies provide us with conflicting results?

One possible reason for conflicting results is that one or more of the experiments was poorly conducted — certain principles of sound experimentation may have been violated. Perhaps the most common error in experimentation is the failure to control important extraneous variables. To demonstrate

briefly how such a failure may produce conflicting results, let us assume that one important extraneous variable is not considered by the experimenter. Unknown to the experimenter, this variable is actually influencing the dependent variable. In one experiment it happens to assume one value, while in a second experiment on the same problem, it happens to assume a different value. Thus, it leads to different values for the dependent variable in the two experiments. The publication of two independent experiments with conflicting conclusions then presents the psychological world with a problem. The solution to this particular problem can be achieved by systematically varying the extraneous variable in a repetition of the two experiments. Let us illustrate by considering a set of experiments having to do with language suppression. In one experiment (Webster and Weingold, 1965) two pronouns were selected and repeatedly exposed in a variety of sentences to the subjects of an experimental group. Control subjects were exposed to the same sentences except that other pronouns were substituted for the special two. From a larger list of pronouns (that contained those two of special interest) both groups of subjects selected a pronoun to use in a sentence. More specifically, the subjects were told to compose sentences using any of the pronouns from the list. It was found that the experimental subjects tended to avoid one of those pronouns to which they had previously been exposed, relative to the frequency of their selection by the control group. It was concluded that prior verbal stimulation produces a satiation effect so that there is a suppression of pronoun choice. Regardless, however, of the theoretical issues involved, our purpose is satisfied by focusing on one of the extraneous variables in the experiment, viz., the location of the experimenter who presented the verbal materials. In this experiment the experimenter sat outside the view of the subject. Albrecht (1965) undertook to repeat the experiment, though she sat in a position that allowed the subjects to see her; quite possibly the subjects could thus receive additional cues, such as those that occurred when the experimenter recorded response information. The results of this repetition, needless to say, did not show a suppression effect of the two pronouns by the experimental group. Not to be discouraged, however, Albrecht and Webster (1966) again repeated the original experiment except that this time it was made certain that the experimenter was hidden from the subject's view. And this time the results confirmed Webster and Weingold's original findings.

It would thus appear that the extraneous variable of experimenter location was sufficiently powerful to influence the dependent variable scores. The fact that it was different in the second experiment led to results that conflicted with those of the first experiment, thus creating a problem. The problem was apparently solved, however, by controlling this variable, and thus the reason for the conflicting results was established. We may only add that it

would have been preferable to have repeated the first two experiments simultaneously, in place of the third, systematically varying experimenter location by means of a factorial design. But this point will be delayed until we take up the factorial design later in the book.

EXPLAINING A FACT

The third way in which we become aware of a problem is when we are in possession of a "fact," and we ask ourselves "Why is this so?" A fact, existing in isolation from the rest of our knowledge, demands explanation.

A science consists not only of knowledge, but of systematized knowledge. The greater the systematization, the greater is the scientist's understanding of nature. Thus, when a new fact is acquired, the scientist seeks to relate it to the already existing body of knowledge. But he does not know exactly where in his framework of knowledge the new fact fits, or even that it will fit. If, after sufficient reflection, he is able to appropriately relate the new fact to existing knowledge, it may be said that he has *explained* it. That fact presents no further problem. On the other hand, if the fact does not fit in with existing knowledge, a problem is made apparent. The collection of additional information is necessary so that eventually, the scientist hopes, the new fact will be related to additional information in such a manner that it will be "explained." By this gradual process, the scientist's understanding and control of nature is extended. Some problems of this kind will lead to little that is of significance for science, while others may result in major discoveries. Examples of new portions of knowledge that have had revolutionary significance are rare in psychology since it is such a new science, but they are relatively frequent in other sciences.¹ To illustrate how the discovery of a new fact created a problem, the solution of which has had important consequences, consider the following example.

One day the Frenchman, Henri Becquerel, found that a piece of photographic film had been fogged. He could not immediately explain this, but in thinking about it he noticed that a piece of uranium had been placed near the film before the fogging. Existing theory did not suggest that there was any connection between the uranium and the fogged film. But Becquerel suggested that the two events were connected to each other. To relate the events more specifically, he had to postulate that the uranium gave off some unique kind of energy. Working along these lines, he eventually determined that the metal gave off radioactive energy which caused the fogging, for which finding he received the Nobel Prize. This discovery led to a whole

¹Wertheimer's attempts to explain the phi phenomenon may be one such case in psychology.

series of developments that have resulted in present-day theories of radioactivity.

The explanation of a fact constitutes a hypothesis (or theory), and it is characteristic of hypotheses that they also apply to other phenomena. That is, most hypotheses are sufficiently general that they are possible explanations of several facts. Hence, the development of a hypothesis that accounts for one fact may be a fertile source of additional problems in the sense that one may ask: "What other phenomena can it explain?" One of the most engaging aspects of the scientific enterprise is to tease out the implications of a general hypothesis and to subject those implications to additional empirical test. An illustration is Hull's (1943) principles of inhibition. To oversimplify the matter, Hull was presented with the fact of spontaneous recovery — that with the passage of time a response that had been extinguished will recover some of its strength and will again be evoked by a conditioned stimulus. To explain this fact Hull postulated that there is a temporary inhibition factor that is built up each time an organism makes a response. He called this factor "reactive inhibition" and held that it is a tendency to *not* make a response, quite analogous to fatigue. When the amount of inhibition is sufficient in quantity, the tendency to *not* respond is sufficiently great that the response is extinguished. But with the passage of time, reactive inhibition (being temporary, like fatigue) dissipates, and the tendency to not respond is reduced. Hence, the strength of the response increases and it thus can reoccur — the response "spontaneously recovers."

Our point is not, of course, concerned with the truth or falsity of Hull's inhibitory principles, but merely to show that a hypothesis that can explain one behavioral phenomenon can be tentatively advanced as an explanation of other phenomena. For example, the principle of reactive inhibition has also been extended to explain why distributed practice is superior to massed practice, to account for an observed superiority of a whole method over a part method of learning, and so forth (e.g., Calvin, *et al.*, 1961, Chapter XVI). Each such attempt to apply this principle to other phenomena has constituted an additional problem. This, as well as others of Hull's principles, has been extremely fruitful in generating new problems that are susceptible to experimental attack.

We can thus see that the growth of our knowledge progresses as we acquire a bit of information, as we advance tentative explanations of that information, and as we explore the consequences of those explanations. In terms of problems, science is a mushrooming affair. As Homer Dubs (1930) correctly noted, every increase in our knowledge results in a greater increase in the number of our problems. We can, therefore, judge a science's maturity by the number of problems that it has; the more problems that a given science faces, the more advanced it is.

A PROBLEM MUST BE SOLVABLE

Not all questions that people are capable of asking can be answered by science.² As noted in Chapter One, a problem can qualify as the object of scientific inquiry only if it is solvable, as distinguished from an unsolvable problem. And since a hypothesis is a tentative solution of a problem, science deals only with hypotheses that are *testable*. One of the most important activities related to science, and also one of the most complex, is the determination of a criterion of testability. Such a criterion should enable us to determine whether or not a problem is solvable, i.e., a problem is solvable if, and only if, it is possible to advance a testable hypothesis as a tentative solution of it. Because of the numerous discordant views on testability, it would be impossible to offer a presentation that would satisfy all. As a result we shall get into this matter only insofar as it affects the everyday life of the experimental psychologist.³

THE TRUTH THEORY OF TESTABILITY

Briefly, we shall say that a problem is solvable if there can be empirical reasons for answering it in a "yes" or "no" fashion. There are two important stages in the development of the theory of testability that we are following here. The first stage is the statement of a truth (or verifiability) theory of testability. The second stage, an improvement on the truth theory, is the statement of a probability theory of testability.

The main principle of the truth theory with which we shall be concerned may be expressed as follows: *A proposition is testable if, and only if, it is possible to determine that the proposition is either true or false.* It follows that only a proposition (a statement or sentence) can be testable. Hypotheses are propositions, so we can use the truth theory to determine whether or not hypotheses are testable. For if it is possible to determine that a hypothesis is true or false, then the hypothesis is testable. But if it is not possible to determine that the

²To be more precise, if a "question" can't be answered, it's really not a question at all. Hence, what might appear to be questions or problems merely because of their grammatical construction are more appropriately called "pseudoquestions" or "pseudoproblems." An example that nobody would waste his time on is "Is he either and/or I during somebody?" but much effort has been expended on pseudoquestions of the sort "Does Ra (the Egyptian Sun God) have a green beard?"

³Historically, the problem has often been approached by the use of the terms *meaning* and *meaningfulness*. We shall here prefer the more neutral words *solvable* when referring to a problem and *testable* when referring to a hypothesis. For more advanced treatments of the topic refer to Reichenbach (1938), Frank (1956), Feigl and Scriven (1956), especially the chapter in the latter by Rudolph Carnap entitled "The Methodological Character of Theoretical Concepts," and Hempel (1965). We shall here largely follow Reichenbach's development.

proposition is either true or false, then the hypothesis is not testable and should be discarded as being worthless to science.

Second, it follows that knowledge can only be expressed in the form of propositions. Problems are best stated in the form of questions, and such questions must be answerable if they are to be subjected to scientific inquiry. When we say that a problem (stated as a question) is solvable, it must be possible to state a hypothesis as a potential answer to the problem, and it must be possible to determine that the hypothesis is either true or false. In short, a solvable problem is one for which a hypothesis that is testable by the truth criterion can be stated.⁴

THE PROBABILITY THEORY OF TESTABILITY

The words "true" and "false" have been used frequently in the above discussion. Strictly speaking, these words have been used only as approximations, for it is impossible to determine beyond all doubt that any given empirical proposition is true or false. The kind of world that we have been given for study is simply not of this nature. The best that we can do is to say that a certain proposition has a determinable degree of probability.⁵ Thus, we cannot say in a strict sense that a certain proposition is true, but the best that we can say is that it is probably true. Similarly, we cannot say that another proposition is false; rather, we must say that it is highly improbable. Strictly speaking, then, the truth theory is inadequate for our purposes, for according to it, no empirical proposition would ever be known to be testable since no empirical proposition can ever be (absolutely) true or false. Hence we shall substitute the probability theory for the truth theory, the essential difference between the two being that the words "a degree of probability" are substituted for "true" and "false." The main principle of the probability theory with which we shall be concerned is: A proposition is testable if, and only if, it is possible to determine a degree of probability for it.

When we say that a proposition is testable, then, we understand that it is testable as defined by the probability theory of testability. In short a problem is solvable if: (1) a relevant hypothesis can be advanced as a tentative solution for it and (2) it is possible to test that hypothesis by determining a degree of probability for it.

⁴Of course the hypothesis must be *relevant* to the problem. For instance, if our problem is "What is the average height of Pigmyes?" an irrelevant (but probably true) hypothesis would be "If a person smokes opium, then he will develop hallucinations." By relevance we shall mean that an inference can be made from the hypothesis to the problem, and the results of that inference constitute a solution to the problem.

⁵By "degree of probability" we mean that the proposition is true with a probability somewhere between 0 (absolutely false) and 1 (absolutely true). For instance, if a proposition has a probability of .5, then it is just as likely to be true as false.

KINDS OF POSSIBILITIES

Let us now enlarge our understanding of the probability theory of testability. In particular, let us focus on the word "possible" contained in our statement of it. What does "possible" refer to? Does it mean that we can test the hypothesis *now*, or at some time in the future? Consider the question "Is it possible for man to fly to Uranus?" If by "possible" here we mean that one can step into a rocket ship today and set out on a successful journey, then clearly such a venture is not possible. But if we mean that such a trip is likely to be possible sometime in the future, then the answer is in the affirmative. In the following simplified discussion we shall consider two interpretations of "possible." The first interpretation we shall call *presently attainable*, and the second *potentially attainable*.

Presently attainable. This interpretation of "possible" concerns those possibilities that lie within the powers of people at the present time. If a certain task can be accomplished with the equipment that is immediately available, we would say that the solution to the task is presently attainable. But if the task cannot be accomplished with tools that are presently available, the solution is not presently attainable. For example, building a bridge over the Suwannee River is presently attainable, but living successfully on Venus is not presently attainable.

Potentially attainable. This interpretation concerns those possibilities that *may* come within the powers of people at some future time, but which are not possessed at the present. Whether or not they actually will be possessed in the future is a difficult matter to decide now. In the event that technological advances are sufficiently successful that we actually come to possess the powers, then the potentially attainable becomes presently attainable. For example, a trip to Uranus is not presently attainable, but we fully expect such a trip to be technologically feasible in the future. Successful accomplishment of such a venture is "proof" that the task should be shifted into the presently attainable category. Less stringently, when we can specify the procedures for solving a problem, and when it has been demonstrated that those procedures can actually be used, then we may shift the problem from the potentially to the presently attainable category.

CLASSES OF TESTABILITY

With these two interpretations of the word "possible" in hand we may now consider two classes of testability, each based on our two interpretations.

Presently testable. If the determination of a degree of probability for a proposition is presently attainable, then the proposition is presently testable. This statement allows considerable latitude, which we must have in order to justify work on problems which have a low probability of being satisfactorily

solved as well as on straightforward, cut-and-dried problems. If one can conduct an experiment in which the probability of a hypothesis can be ascertained with the tools that are presently at hand, then clearly the hypothesis is presently testable. If we cannot now conduct such an experiment, the hypothesis is not presently testable.

Potentially testable. A proposition is potentially testable if it may be possible to determine a degree of probability for it at some time in the future, if the degree of probability is potentially attainable. While such a proposition is not presently testable, improvement in our techniques and the invention of new ones may make it possible to test it later. Within this category we also want to allow wide latitude. There may be statements for which we know with a high degree of certainty how we will eventually test them, though we simply cannot do it now. At the other extreme are statements for which we have a good deal of trouble imagining the procedures by which they will eventually be tested, but we are not ready to say that someone will not some day design the appropriate tools.

A WORKING PRINCIPLE FOR THE EXPERIMENTER

On the basis of the above considerations, we may now formulate our principles of action with regard to hypotheses. First, since the psychologist conducting experiments must work only on problems that have a possibility of being solved with the tools that are immediately available, he must apply the criterion of present testability in his everyday work. Therefore, only if it is clear that a hypothesis is presently testable should it be considered for experimentation. The psychologist's problems which are not presently but are potentially testable, should be set aside in a "wait and see" category. When sufficient advances have been made so that the problem can be investigated with the tools of science, it becomes presently testable and can be solved. If sufficient technological advances are not made, then the problem is maintained in the category of potential testability. On the other hand, if advances show that the problem that is set aside proves not to be potentially testable, it should be discarded as soon as this becomes evident, for no matter how much science advances, no solution will be forthcoming.

APPLYING THE CRITERION OF TESTABILITY

If we examine the ways in which problems are stated by students and by scientists, we can specify three reasons why they may be unsolvable, and hence why corresponding hypotheses may be untestable. The first is that the

problem may be so unstructured and vaguely stated that we wouldn't even know how to commence solving it. The second reason is that certain terms in the question may be unclear, ambiguous, or ill-defined. And third, while the question may be sufficiently precise and we can adequately define the terms, there may be simply no way to ascertain a degree of probability for the proposed solution. Since the proper formulation of a problem is basic to the conduct of an experiment, it is essential that the experimenter avoid these pitfalls. Let us, then, explore these three matters in greater detail.

THE UNSTRUCTURED PROBLEM

By this stage in your study of psychology, your sophistication is such that this type of problem would not occur to you, and the category is offered here merely to enhance your perspective. Your college instructor, however, must face the fact that the vague, inadequately formulated problem will be presented to him by his introductory students for many generations to come. How, for instance, can he answer such questions as "What's the matter with his (my, your) mind?" "How does the mind work?" "Is it possible to change human nature?" and so forth. These problems are unsolvable because the intent is unclear and the domain to which these problems refer is so amorphous that it is impossible to specify what the relevant observations would be. After lengthy discussion with the asker, however, it *might* be possible to determine what he is trying to ask and to thereby reformulate the question so that it does become answerable. Perhaps, for example, suitable dissection of the question "What's the matter with my mind?" might lead to a reformulation such as "Why am I compelled to count the number of door knobs in every room that I enter?" Such a question is still difficult to answer, but at least the chances of success are increased because the question is now more precisely stated and refers to events that are more readily observable. Whether the game is worth the candle is another matter. For the personal education of the student, it probably is. Reformulations of this type of question, however, are not very likely to advance science.

INADEQUATE DEFINITIONS

Of course the vaguely stated problem typically contains terms that are inadequately defined, which contribute to its vagueness. But there may be problems that are solvable if we but knew what was meant by one of the terms contained in their statement. Consider, for example, the topical question, "Can machines think?" This is the contemporary analogue of the question that Thorndike took up in great detail early in the century, viz., "Do lower animals reason?" Whether or not these problems are solvable depends on how "think" and "reason" are defined. Unfortunately, much

energy has been expended in arguing such questions in the absence of clear specifications of what is meant by the crucial terms. Historically, the disagreements between the disciples, Jung and Adler, and the teacher, Freud, are a prime example. Just what *is* the basic driving force? Is it the libido, with a primary emphasis on sexual needs? Is it Jung's more generalized concept of the libido as "any psychic energy?" Or is it, as Adler held, a compensatory energy, a "will to power?" This problem, it is safe to say, will continue to go unsolved until these hypothesized concepts are adequately defined, if in fact they ever are.

A question that is receiving an increasing amount of attention from many points of view is "How do children learn language?" In their step by step accounts of the process, linguists and psychologists frequently include a phase in language development which may be summarized as "Children then learn to imitate the language production of adults around them." The matter is usually left there with the feeling that this highly complex process is well accounted for. A closer analysis of "Do children learn language by imitation?" however, leads us to be not so hasty. Because we don't know what the theorist means by "imitation" — its sense may vary from a highly mystical interpretation to a concrete, objectively observable behavioral process — the question is unsolvable at this state of its formulation. ✓

One of the main reasons that such problems as those considered above are unsolvable as they have been stated is that many of the terms have been imported from everyday language. Our common language is replete with ambiguities, as well as with multiple definitions for any given word. If one does not give cognizance to this point, he can expend his argumentative (and research) energies in vain. Everyone can recall, no doubt, at least several lengthy and perhaps heated arguments which, on more sober reflection, were found to have resulted from a lack of agreement on the definition of certain terms that were basic to the discussion. To illustrate, suppose a group of people carried on a discussion about happiness. The discussion would no doubt take many turns, produce many disagreements, and probably result in considerable unhappiness on the part of the disputants. It would probably accomplish little, unless at some early stage in the discussion the people involved were able to agree on an unambiguous definition of "happiness." Although it is impossible to guarantee the success of a discussion in which the terms are adequately defined, without such an agreement there would be no chance of success whatsoever.

✓ The importance of adequate definitions in science cannot be too strongly emphasized. The main functions of good definitions are: (1) to clarify the phenomenon under investigation; and (2) to allow us to communicate with each other in an unambiguous manner. These functions are accomplished by operationally defining the empirical terms with which the scientist deals.

When one faces the problem of how to define a term operationally he, in

large part, addresses himself to the question of whether or not his problem is solvable. Hence, if the terms contained in his problem can be operationally defined, then the scientist has moved a long way toward rendering the problem solvable. If, to follow the examples presented above, terms such as "thinking" or "imitation" can be operationally defined, then the questions which include them may be classified as solvable. Otherwise, they are only potentially solvable or unsolvable.

Essentially, an operational definition is one that indicates that a certain phenomenon exists and does so by specifying precisely how (and preferably in what units) the phenomenon is measured. That is, an operational definition of a concept consists of a statement of the operations necessary to produce the phenomenon. Once the method of recording and measuring a phenomenon is specified, that phenomenon is said to be operationally defined. The precise specification of the defining operations obviously accomplishes the intent of the scientist — by performing those operations, a phenomenon is produced and a number of observers can agree on the existence and characteristics of the phenomenon. Hence, a phenomenon that is operationally defined is reproducible by other people. Because we operationally define a concept, the definition of the concept consists of the objectively stated operations performed in producing it. Others can then reproduce the phenomenon by repeating these operations. For example, when we define air temperature, we mean that the column of mercury in a thermometer rests at a certain point on the scale of degrees. Consider the psychological concept of hunger drive. One way of operationally defining this concept is in terms of the amount of time that an organism is deprived of food. Thus, one operational definition of *hunger drive* would be a statement about the number of hours of food deprivation. Accordingly, we might say that an organism that has not eaten for 12 hours is more hungry than one that has not eaten for two hours.

A considerable amount of work has been done in psychology on steadiness. There are a number of different ways of measuring steadiness, and accordingly there would be a number of different operational definitions of the concept. For example, let us consider the Whipple Steadiness Test. The apparatus in this test consists of a series of holes, varying from large to small in size, and a stylus. The subject attempts to hold the stylus as steady as he can in each hole, one at a time, without touching the sides. The number of contacts he makes is automatically recorded. Presumably, the steadier subject will make fewer contacts than the unsteady subject. Therefore, we would operationally define steadiness as the number of contacts made by a subject when taking the Whipple Steadiness Test. Let us add that if we measured steadiness by using other types of apparatus, then we would have additional operational definitions of steadiness.

We can now see that the first step in solving a problem is to ask whether or not the critical empirical terms can be operationally defined. What we

are basically requiring is a specification of the laboratory methods and techniques for producing stimulus events and for recording and measuring response phenomena. We must be able to refer to (or "point" to) some event in the environment that corresponds to each empirical term in the statement of problems and hypotheses. If no such operation is possible for all of these terms, we must conclude that the problem is unsolvable and that the hypothesis is untestable. In short, by subjecting the problem to the criterion of operational definition of its terms, we render a solvable-unsolvable decision, on the basis of which we either continue or abandon our research on that question.⁶

IMPOSSIBILITY OF OBTAINING RELEVANT DATA

Although there is some overlap among the three reasons that problems are unsolvable, the emphasis in each case differs. In this category the emphasis is on the type of question that is sufficiently precise and whose terms are operationally definable, but we are at a loss to specify how we would collect the necessary data. As an illustration, consider the question of the effect of psychotherapy on the intelligence of a clinical patient who cannot talk. Note that we can adequately define the crucial terms such as intelligence and therapy. The patient, we observe, scores low on an intelligence test. After considerable clinical work his speech is improved; he is given another intelligence test and registers a significantly higher score. Did the intelligence of the patient actually increase as a result of the clinical work? Alternatives are possible: Was the first intelligence score invalid because of the difficulties of administering the test to the non-verbal patient? Did the higher score result from merely "paying attention" to the person? Was he going through some sort of transition period such that merely the passage of time (with various experiences) gave him the opportunity to increase his score? Clearly it is impossible to decide among these possibilities and the problem is unsolvable as stated.

"If you attach the optic nerve to the auditory areas of the brain, will you sense visions auditorily?" Students will probably continue arguing this

⁶In 1927 P. W. Bridgeman's book was published, initiating the movement known as operationism. The prime assumption of operationism is that the adequate definition of the variables with which a science deals is a prerequisite to advancement. In the years that have followed much has been written concerning operationism; these writings have led to many polemics. We have chosen to avoid becoming involved in such problems, but it is worthwhile to point out that they exist. The student should thus realize that a discussion of operational definitions can lead into matters far beyond what we have taken up here. For some of the more philosophical discussions of this topic consult Frank (1956), while for a more extended consideration of how operational definitions are used in psychology read the excellent discussion in Underwood (1957).

question until neurophysiological technology progresses to the point that we can change this potentially solvable problem to the presently solvable category.

A similar candidate for dismissal from the presently solvable category is a particular attempt to explain reminiscence, a phenomenon that may appear under certain conditions. To illustrate briefly, let us say that a person practices a task such as memorizing a list of words, though the learning is not perfect. He is tested immediately after a certain number of practice trials. Then, after a period of time during which he has no further practice, he is tested on the list of words again. On this second test it may be found that he recalls more of the words than he did on the first test. This is *reminiscence*. We may say that reminiscence occurs when the recall of an incompletely learned task is more complete after a period of time has elapsed following the learning period than it is immediately after the learning period. The problem is how to explain this phenomenon.

One possible explanation of reminiscence is that, while there are no formal practice trials following the initial learning period, the subject continues implicitly to practice the task. That is, he "rehearses" the task "to himself," following the initial practice period, and before the second test. This informal rehearsal could well lead to a higher score on the second test. Now our purpose is not to take issue with this attempt to explain reminiscence. Rather, we wish to examine a line of reasoning that led one psychologist to reject "rehearsal" as an explanation of the phenomenon.

The psychologist lists several types of evidence which suggest that rehearsal cannot account for reminiscence. Among this evidence is the following statement: "Rats show reminiscence in maze learning (Bunch and Magsdick, 1933), and it is not easy to imagine rats rehearsing their paths through a maze between trials" (Deese, 1952, p. 175). Such a statement cannot seriously be considered as bearing on the problem of reminiscence — there is simply no way at present to determine whether rats do or do not rehearse, assuming the common definition of "rehearse." Hence, the hypothesis that rats show reminiscence but do not rehearse is not presently testable. (This does not mean of course, that the rehearsal hypothesis or other explanations of reminiscence are untestable).

To take up another example of an unsolvable problem, let us consider testing two theories of forgetting: the *disuse theory*, which says that forgetting occurs strictly because of the passage of time; and the *interference theory*, which says that forgetting is the result of competition from other learned material. Which theory is more probably true? An experiment by Jenkins and Dallenbach (1924) is frequently cited as evidence in favor of the interference theory, and this is scientifically acceptable evidence. This experiment showed that there is less forgetting during sleep (when there is little interference) than

during waking hours. However, their data indicate considerable forgetting during sleep, which is usually accounted for by saying that even during sleep there is some interference (possibly from incoming stimuli, dreaming, etc.). To determine whether or not this is so, to test the theory of disuse strictly, we must have a condition in which a person has zero interference. Technically, there would seem to be only one condition which might satisfy this requirement — death — and even this is doubtful for it would be difficult to measure retention under such a condition. Thus, the Jenkins-Dallenbach experiment does not provide a completely general test of the theory of disuse. Therefore, we must consider the problem of whether, during a condition of zero interference there is no forgetting, as a presently unsolvable problem, though it is potentially solvable (perhaps by freezing the subjects).

The interested student should consider for himself a number of other problems in psychology to determine whether they are solvable. For example: "Does a subject in an experiment perform exactly as he would if he were not in an experiment?" Can we answer the question of how the subject would perform if the apparatus or the questionnaire or the test were not used?

Before concluding this section it may be instructive to take up a particular kind of reasoning that, when it occurs, is outrightly disastrous for the scientific enterprise. This fallacious reasoning, called *vicious circularity*, occurs when an answer is based on a question and the question on the answer, with no appeal to other information outside of the vicious circle. A historical illustration is the development and demise of the instinct doctrine. In the early part of our century "instinct naming" was a very popular game, and it resulted in quite a lengthy list of such instincts as gregariousness, pugnacity, etc. The goal was to explain the occurrence of a certain kind of behavior, call it X, by postulating the existence of an instinct, say Instinct Y. Only eventually did it become apparent that this endeavor led exactly nowhere, at which time it was discontinued. The game, to reconstruct its vicious circularity, went thusly — Question: "Why do organisms exhibit behavior X?" Answer: "Because they have instinct Y." But the second question: "How do we know that organisms have Instinct Y?" Answer: "Because they exhibit behavior X." The reasoning goes from X to Y and from Y to X, thus explaining nothing. Problems that are approached in this manner constitute a unique class of unsolvable ones, and we must be careful to avoid the invention of new games such as "drive naming." Let us illustrate the danger from a more contemporary point of view. One question that may be important in certain contexts is why a given response did not occur. A possible answer is that an inhibitory neural impulse prevented the excitatory impulse from producing a response. That is, recent neurophysiological research has indicated the existence of efferent neural impulses that descend from the central nervous system, and they may inhibit responses. The behaviorist, if he relies on this

concept, may fall into a trap similar to that of the instinct doctrinists, i.e., to answer the question "Why did response X fail to occur?" he could answer: "Because there was an inhibitory neural impulse." Whereupon we must ask the second question again: "But how do you know that there was an inhibitory neural impulse?" and if he answers, in effect, "Because the response failed to occur," we can immediately see that the process of vicious circularity has been involved. To avoid this fallacious reasoning, the psychologist must rely on outside information. In this instance, he should independently record the inhibitory neural impulse, so that he has a sound, rather than a circular, basis for asserting that it occurred. Hence, the reasoning could legitimately go as follows: "Why did response X fail to occur?" "Because there was a neural impulse that inhibited it." "How do we know that there actually was such an impulse?" "Because we recorded it by a set of separate instruments," as did Hernandez-Peon, Scherrer, and Jouvet (1956).

The lesson from these considerations of vicious circularity is, to be brief, that there must be documentation of the existence of phenomena that is independent of the statement of the problem and its proposed solution. Otherwise the problem is unsolvable — there is no alternative to the hypothesis than that it be true. Guthrie's principle of learning states that "A combination of stimuli which has accompanied a movement will on its recurrence tend to be followed by that movement" (1952, p. 23). Stated more simply, when a response is once made to a stimulus pattern, the next time the stimulus pattern is presented, the organism will make the same response. To test his principle, suppose that we record a certain response to the stimulus. Then we later present the stimulus and find that a different response occurs. One might conclude that this finding disconfirms Guthrie's principle. Or, if the scientist falls victim to the vicious circularity line of reasoning, he might say that, while the second presentation of the stimulus appeared to be the same as the first, it must not have been. Because the response changed, the stimulus, in spite of efforts to hold it constant, must have changed in some way that was not readily apparent. A scientist who reasons thusly would never be able to falsify the principle, and hence the principle becomes untestable. To render the principle testable there must be a specification of whether or not the stimulus pattern changed from the first to the second test of it that is independent of the response findings.

SOME ADDITIONAL CONSIDERATIONS OF PROBLEMS

Even after we have determined that a problem is presently testable, there are other requirements to be met before considerable effort is expended in conducting an experiment. One such requirement is that the problem be

sufficiently important. Numerous problems arise for which the psychologist will furnish no answers immediately or even in the future, though they are in fact solvable problems. Some problems are just not important enough to justify research — they are either too trivial, or too expensive (in terms of time, effort and money) to answer. For instance, the problem of whether rats prefer Swiss or American cheese is likely to go unanswered for centuries; similarly, “why men fight” not because it is unimportant, but because its answer would require much more effort than society seems willing to expend on it.

Sometimes the psychologist is aware of a problem that is solvable, adequately phrased, and important, but an accumulation of experiments on the problem shows contradictory results. And often there seems to be no reason for such discrepancies. That is what might be called “the impasse problem.” When faced with this situation, it would not seem worthwhile to conduct “just another experiment” on the problem, for little is likely to be gained, regardless of how the experiment turns out. That is, if the experiments are numerous and contradictory, little is to be gained by adding one more set of data to either side. Unless the experimenter can be extremely imaginative and develop a totally new approach that has some chance of systematizing the knowledge in the area, he probably should stay out of it and use his energy to perform research on a problem that has a greater chance of contributing some new knowledge.

Some aspects of this general discussion may strike you as representing a “dangerous” point of view. One might ask how we can ever know that a particular problem is really unimportant. Perhaps the results of an experiment on what some regard as an unimportant problem might turn out to be very important — if not today, perhaps in the future. Unfortunately, there is no answer to such a position. Such a situation is, indeed, conceivable, and our position as stated above would “choke off” some important research. It is suggested, however, that if an experimenter can foresee that his experiment will have some significance for theory or for some applied practice, his results are going to be more valuable than if he cannot foresee such consequences. There are some psychologists who would never run an experiment unless it has some specific influence on a given theoretical position. This might be too rigid a position, but it does have merit.

There is no clear delineation between an important problem and an unimportant one, but it can be fairly clearly established that some problems are more likely to contribute to the advancement of psychology than are others. And it is a good idea for the experimenter to try to choose what he considers an important problem rather than a relatively unimportant problem. Within these rather general limits, no further restrictions are suggested. In any event, science is the epitome of the democratic process, and any scientist is

free to work on whatever problem he chooses. What some scientists would judge to be "ridiculous problems" may well turn out to have revolutionary significance. Some psychologists have wished for the creation of a professional journal with a title like "The Journal of Crazy Ideas," to encourage wild and speculative research.

PSYCHOLOGICAL REACTIONS TO PROBLEMS

Unfortunately the existence of problems that lead to scientific advances are frequently a source of anxiety for some people. When there is a new discovery, people tend to react in one of two ways. The curious, creative person will adventurously attempt to explain it. The incurious and unimaginative person, on the other hand, may attempt to ignore the problem, hoping it will "go away." A good example of the latter type of reaction occurred around the 15th century when mathematicians produced a "new" number they called "zero." The thought that zero could be a number was disturbing and a number of city legislative bodies even passed laws forbidding its use. The creation of imaginary numbers led to similar reactions; in some cases the entire arabic system of numerals was outlawed. Fortunately, this legislation was not effective.

Strangely, negative reactions to scientific discoveries have not been confined to the layman; in fact they have frequently been quite pronounced and emotional on the part of outstanding scientists. A fascinating historical consideration of factors that limit the openmindedness of scientists has been presented by Barber (1961). He discusses, for example, the reasons that it took astronomer-scientists so long to accept the Copernican theory of planetary motion. Among the reasons for rejecting the Copernican theory was that it is "simply absurd," to think that the earth moves; this is hardly a basis for scientific reasoning. Similarly, Mendel's great achievement — the development of his theory of genetic inheritance — failed to be accepted because it ran counter to the existent theories of the time, because it was "too mathematical," and so forth. And, merely because English astronomers of 1845 distrusted mathematics, Adams' discovery of a new planet (Neptune) was not published.

One major error that has been committed by scientists throughout history, even up to the present time, is judging the quality of scientific research by the status of the researcher. The most interesting problems that Mendel faced in this respect are worth exploring a bit further. Mendel, it seems, wrote deferentially to one of the distinguished botanists of the time, Carl von Nägeli of Munich. This unimportant monk from Brünun was obviously a mere amateur expressing fantastic notions that ran, incidentally, counter to

those of the master. Nevertheless, von Nägeli honored Mendel by answering him and by advising him to change from experiments on peas to hawkweed. It is ironic, indeed, that Mendel took the advice of the great man and thus labored in a blind alley for the rest of his scientific life on a plant not at all suitable for the study of inheritance of separate characteristics.

It is to be hoped that society in general, and scientists in particular, will eventually learn to assess advances in knowledge on the basis of a truth criterion alone, and that the numerous sources of resistance to discoveries will be reduced and eliminated.

THE HYPOTHESIS

THE NATURE OF A HYPOTHESIS

We have previously noted that a scientific investigation starts with the statement of a solvable problem. Following this, a tentative solution to that problem is offered in the form of a proposition. The proposition must be testable — it must be possible to determine whether it is probably true or false. Thus, a hypothesis is a testable proposition that *may be* the solution of a problem. To enlarge on this definition of a hypothesis briefly consider the relationship between the problem and the hypothesis.

First, if it is found after suitable experimentation that the relevant hypothesis is probably true, then we may say that the hypothesis *solves* the problem to which it is addressed. If the relevant hypothesis is probably false, we may say that it does not solve the problem. To illustrate, let us consider the problem: “Who makes a good bridge player?” Our hypothesis might be that “people who are intelligent and who show a strong interest in bridge make good bridge players.” The collection and interpretation of sufficient data

might confirm the hypothesis. In this case we say that we have solved the problem because we can answer the question.¹

On the other hand, let us say that we fail to confirm our hypothesis. In this event it should be apparent that we have not solved the problem, i.e., we have failed to obtain definite information that these specific qualities make a good bridge player.

Frequently when we obtain a true hypothesis and thus solve a problem we may say that the hypothesis *explains* the phenomena with which the problem is concerned. Let us assume that a problem exists because we are in possession of a certain fact. As noted in the last chapter, this fact, existing in isolation, requires an explanation. The need to explain the fact presents us with our problem. If we can relate that fact to some other fact in an appropriate manner, we can say that the first fact is explained. A hypothesis is the tool by which we seek to accomplish such an explanation. That is, we use a hypothesis to state a possible relationship between one fact and another. If we find that the two facts are actually related in the manner stated by the hypothesis, then we have accomplished our immediate purpose — we have explained the first fact. (A more complete discussion of explanation is offered in Chapter Fourteen.)

To illustrate, let us return to an example used in Chapter Two. Becquerel was presented with the fact that a certain photographic film had been fogged. This fact demanded an explanation. In addition to noting this fact, Becquerel also noted a second fact: that a piece of uranium was lying near the photographic film. His hypothesis was that some characteristic of the uranium produced the fogging. Becquerel's test of this hypothesis proved successful; he established that his hypothesis was true. Thus, by relating the fogging of the film to a characteristic of the uranium he explained his fact.

But what do we mean by "fact"? "Fact" is a common-sense word, and as such its meaning is rather vague. We understand something by it, such as a fact is "an event of actual occurrence." It is something that we are quite sure has happened (Becquerel was quite sure that the photographic film was fogged). But it may facilitate our task if we replace this common-sense word with a more precise term. For instance, instead of using the word *fact*, suppose that we conceive of the fogging of the film as a *variable*, that is, the film may be fogged in varying degrees, from a zero amount up to some large amount, such as total exposure. Similarly, the amount of radioactive energy that is given off by a piece of uranium may be conceived of as a variable; it may be an amount anywhere from zero to a large amount. Therefore, instead of

¹The only qualification is that the problem is not *completely* solved. We are using "solved" in an approximate sense and further research is required to enlarge our solution, as for instance, finding other factors that make good bridge players. In this case we may eventually arrive at a hypothesis that has a greater probability than the earlier one; and it offers a more complete solution.

saying that two *facts* are related, we may now make the more desirable statement that two *variables* are related. The advantages of this procedure are sizeable. For example, we may now hypothesize a *quantitative* relationship — the greater the amount of radioactive energy given off by the piece of uranium, the greater the fogging of the photographic plate. Hence, instead of making the rather crude distinction between “fogged” and “unfogged” film, we may now talk about the *amount* of fogging. Similarly, the uranium is not simply giving off radioactive energy, it is emitting an *amount* of energy. We are now in a position to make statements of wide generality. Before, we could only say that if the uranium gave off energy, the film would be fogged. Now, we can say, for instance, that if the uranium gives off a “little” energy, the film will be fogged a small amount; if the uranium gives off a lot of energy, the film will be greatly fogged, and so on. Of course, this example is only illustrative; we can make many more statements about the relationship of these two variables with the use of numbers. But this matter will be taken up more thoroughly shortly.

These considerations now allow us to enlarge on our preceding definition of a hypothesis. For now we may define a hypothesis as *a testable statement of a potential relationship between two (or more) variables*.

There are a number of terms used in science other than “hypothesis” to refer to statements of relationships between variables. They are such words as “theories,” “laws,” “principles,” and “generalizations.” The discussion at the present level will be applicable to any statement of an empirical relationship between variables, without here distinguishing among them. The point that we wish to focus on is that an experiment is conducted to test an empirical relationship and, for convenience, we will usually refer to that relationship as a hypothesis. That a hypothesis is *empirical* means that it directly refers to data that we can obtain from our observation of nature. More precisely, the variables contained in an empirical hypothesis are operationally definable and hence refer to events that can be directly observed and measured.

ANALYTIC, CONTRADICTIONARY, AND SYNTHETIC STATEMENTS

This last point above is important for the understanding of the nature of a hypothesis. It will be advantageous to consider it more fully. To accomplish this let us note that all possible statements fall in one of three categories: *analytic*, *contradictory*, or *synthetic*. These three kinds of statements differ on the basis of their possible *truth values*. By *truth value* we mean whether a statement is true or false. Thus, we may say that a given statement has the truth value of *true* (such a statement is “true”) or that it has the truth value of *false* (this one is “false”). Because of the nature of their construction (the way in which

they are formed), however, some statements can take on only certain truth values. Some statements, for instance, can take on the truth value of *true* only. Such statements are called analytic statements (other names for them are "logically true statements" or "tautologies"). Thus, an analytic statement is a statement that is always true — it cannot be false. The statement "If you have a brother, then, either you are older than your brother or you are not older than your brother" is an example of an analytic statement. Such a statement exhausts the possibilities, and since one of the possibilities must be true, the statement itself must be true. A contradictory statement (sometimes also called a "self contradiction" or a "logically false statement"), on the other hand, is one that always assumes a truth value of false. That is, because of the way in which it is constructed, it is necessary that the statement be false. A negation of an analytic statement is obviously a contradictory statement. For example, the statement "it is false that you are older than your brother or you are not older than your brother (or the logically equivalent statement "If you have a brother, then you are older than your brother and you are not older than your brother") is a contradictory statement. Such a statement includes all of the logical possibilities, but says that all of these logical possibilities are false.

The third type of statement is the synthetic statement. A synthetic statement may be defined as any statement that is neither an analytic nor a contradictory statement. In other words, a synthetic statement is one that may be *either* true or false, or more precisely, it has a probability of being true or false. An example of a synthetic statement would be "You are older than your brother," a statement that may be either true or false. And the important point to observe in this discussion is that a *hypothesis must be a synthetic statement*. Thus any hypothesis must be capable of being proven (probably) true or false.

The differences among the three types of statements are highlighted in Table 3.1. There we may see that the symbolic statement of an analytic

Table 3.1. *Possible kinds of statements*

Type of Statement:	Analytic	Synthetic	Contradictory
Symbolic Statement	a or not a	a	a and not a
Example of Statement:	"I am in Chicago, or I am not in Chicago."	"I am in Chicago."	"I am in Chicago and I am not in Chicago."
Truth or Probability Value:*	(Absolutely) true	$1.0 > P > 0$	(Absolutely) false

*The symbol $>$ may be read as "greater than" or $<$ as "less than." Thus, for the synthetic statement, P is less than 1.0 but zero is less than P . Or alternatively, 1.0 is greater than P ; but P is greater than 0.

proposition is " a or not a ." For instance, if we let a stand for the sentence, "I am in Chicago," then the appropriate analytic statement is "I am in Chicago or I am not in Chicago." This statement is necessarily true because no other possibilities exist. The synthetic statement is symbolized by a . Thus, following our example, it says "I am in Chicago," and the probability of this statement is necessarily less than ($<$) 1.0, but necessarily greater than ($>$) 0. The symbolic statement of the contradictory type of proposition is " a and not a " — "I am in Chicago and I am not in Chicago." Clearly, such a statement is absolutely false, barring such unhappy possibilities as being in a severed condition.

Now, why should we state hypotheses in the form of synthetic statements? Why not use analytic statements, in which case it would be guaranteed that our hypotheses are true? The answer to such a question appears in an understanding of the function of the various kinds of statements. The reason that a synthetic statement may be true or false is that it refers to the empirical world, i.e., it is an attempt to tell us something about nature. And as we previously saw, every statement that refers to natural events might be in error. An analytic statement, however, is empty. That is, while it is absolutely true, it tells us nothing about the empirical world. This characteristic results because an analytic statement includes all of the logical possibilities, but it does not attempt to inform us which is the true one. And this is the price that one must pay for absolute truth. If one wishes to state information about nature, he must use a synthetic statement, in which case the statement always runs the risk of being false. Thus, if someone asks me if you are older than your brother I might give him my best judgment, say, "you are older than your brother" which is a synthetic statement. I may be wrong in this statement, but at least I am trying to tell him something about the empirical world. Such is the case with our scientific hypotheses; they may be false in spite of our efforts to assert true hypotheses, but they are potentially informative in the sense that they attempt to say something about nature.

Now, if analytic statements are empty and thus do not tell us anything about nature, why do we bother with them in the first place? The answer to this question could be made quite detailed. Suffice it to say here that analytic statements are quite valuable in facilitating deductive reasoning (logical inferences). The statements in mathematics and logic are analytic and contradictory statements, and are valuable to science because they allow us to transform synthetic statements without adding additional knowledge (recall that analytic statements are empirically empty). The essential point is that sciences use all three types of statements, but use them in different ways. We here emphasize the point that the synthetic type of proposition is used for the statement of hypotheses; for, in stating a hypothesis, we attempt to say something informative about the natural world. This attempt carries with it the possibilities that our hypothesis is probably true or false.

THE MANNER OF STATING HYPOTHESES

Granting, then, that a hypothesis is a statement of a potential empirical relationship between two or more variables, and also that it is possible to determine whether the hypothesis is probably true or false, we might well ask what form that statement should take. That is, precisely how should we state hypotheses in scientific work?

Lord Russell answers this question by proposing that the logical form of the general implication be used for expressing hypotheses (cf. Reichenbach, 1947, p. 356). Using the English language, the general implication may be expressed as: "*If . . . , then . . .*." That is to say, *if* certain conditions hold, *then* certain other conditions should also hold. To better understand the "*If . . . , then . . .*" relationship, let *a* stand for the first set of conditions and *b* for the second set of conditions. In this case the general implication would be "*If a, then b.*" But in order to communicate what the conditions indicated by *a* are, we must make a statement. Therefore, we shall consider that the symbols *a* and *b* are actually statements which express these two sets of conditions. And if we join these two simple statements, as we do when we use the general implication, then we end up with a single compound statement. This compound statement *is* our hypothesis.

The statement *a*, incidentally, is referred to as the *antecedent condition* of the hypothesis (it "comes first"), and *b* is called the *consequent condition* of the hypothesis (it follows the antecedent condition). We have previously noted that a hypothesis is a statement relating two variables. Since we have said that antecedent and consequent conditions of a hypothesis are stated as propositions, it follows that the symbols *a* and *b* are *propositional variables*. A hypothesis, thus, proposes a relationship between two (propositional) variables by means of the general implication as follows: "*If a is true, then b is true.*" The variables *a* and *b* may stand for whatever we wish. If we suspect that two particular variables are related, we might hypothesize a relationship between them. For example, we might think that industrial work groups that are in great inner conflict have decreased production levels. Here the two variables are: (1) the amount of inner conflict in an industrial work group and (2) the amount of production that work groups turn out. We can formulate two sentences; (1) "An industrial work group is in great inner conflict;" and (2) "That work group will have a decreased production level." If we let *a* stand for the first statement and *b* for the second, our hypothesis would read: "*If an industrial work group is in great inner conflict, then, that work group will have a decreased production level.*"

With this understanding of the general implication for stating hypotheses, it is well to inquire about the frequency with which Russell's suggestion has been accepted in psychology. The answer is quite clear: the explicit use of the general implication is almost nonexistent. Two samples of hypotheses,

essentially as they are stated in professional journals, should suffice to illustrate the point:²

1. The present investigation is designed to study the effects of the teacher's praise on reading growth. (Silverman, 1957.)
2. Giving students an opportunity to write comments on objective examinations results in higher test scores (McEachie, Pollie & Speisman, 1955.)

Clearly these hypotheses, or implied hypotheses, fail to conform to the form specified by the general implication. Is this bad? Are we committing serious errors by not precisely heeding Russell's advice? Not really. For as Hempel and Oppenheim (1948) point out, it is always possible to restate such hypotheses as general implications. For example, the above hypotheses could be restated as follows:

The first hypothesis contains two variables: (1) amount of praise and (2) amount of reading growth. The propositions concerned with these variables are: (1) A teacher praises a student for good reading performance, and (2) the student's reading growth increases. The hypothesis relating these two variables may now be expressed. "*If a teacher praises a student for good reading performance, then the student's reading growth will increase.*"

We may similarly identify the variables contained in the second hypothesis and state the propositions as follows: (1) Students are given the opportunity to write comments on objective examination questions, and (2) those students achieve higher test scores. The hypothesis: "*If students are given the opportunity to write comments on objective examination questions, then those students achieve higher test scores.*"

It is apparent that these two hypotheses fit the "*If a then b*" form, although it was necessary to modify somewhat the original statements. Even so, these modifications did not change the nature or the meaning of the hypotheses.

What we have said to this point, then, is that Russell has suggested the use of the general implication for stating hypotheses, and that psychologists do not take his advice in that they express their hypotheses in a variety of ways. However, we can restate their hypotheses as general implications. The next question, logically, is why did Russell offer this advice, and why are we making a point of it here? Briefly, the way in which we determine whether or not a hypothesis is confirmed depends on our making certain inferences from experimental findings to the hypothesis. The rules of logic tell us what kind of inferences we can legitimately make (they are called valid inferences). But in order to determine whether or not the inferences are valid the statements

²The statement of these hypotheses has been modified somewhat for easier comprehension. The interested reader, of course, may consult the original articles.

involved in the inferences (e.g., the hypotheses) must be stated in certain standard forms, one of which is the *general implication*. Hence, in order to discuss the nature of experimental inferences and really to understand them, we must use standard logical forms. This area will be covered more completely in Chapter Twelve, when we consider the nature of experimental inferences.

Another reason is that if the experimenter attempts to state his hypothesis as a general implication it may help him to clarify the prime reason for conducting his experiment. That is, by succinctly and logically writing down the purpose of his experiment as a test of a general implication the experimenter is forced to come to grips with the precise nature of his variables. Any remaining vagueness in the hypothesis can then be removed when the operational definition of the variables is stated.

There is yet another form that is used in the stating of hypotheses. It involves certain mathematical statements that are essentially of the following nature: $Y = f(X)$. That is, a hypothesis stated in this way proposes that some variable, Y , is related to some variable, X , or, alternatively, that Y is a function of X . Such a mathematically stated hypothesis clearly fits our more general definition of a hypothesis, namely that two variables are related. Although the variables in this case are quantitative (their values can be measured with numbers), the two variables may still refer to whatever we wish. For instance, we might refer to hypothesis two on p. 41 and assign numbers to the independent variable. The independent variable, which would be X in the equation $Y = f(X)$, might be stated as the extent to which students are given the opportunity to write comments on examination questions. For instance, we might develop a scale such that one would indicate very little opportunity, two a little greater opportunity, three a medium amount of opportunity, and so on. Test scores, the dependent variable, Y , could be similarly quantified — 100 might be the highest possible score and zero the lowest. Thus, the hypothesis could be tested for all possible numerical values of the independent and the dependent variables.

In any event, the important point here is that even though a hypothesis is stated in a mathematical form, that form is basically of the "If a , then b " relation. Instead of saying "If a , then b " we merely say "If (and only if) X is this value, then Y is that value." For example if X is 3 (a medium opportunity to write comments) then Y is 75 (an average grade).

In concluding this section, let us consider two common misconceptions about the statement of hypotheses as general implications. First, it is erroneous to say that the antecedent conditions *cause* the consequent conditions. This may or may not be the case. The general implication merely states a potential relationship between two variables — *if* one set of conditions holds, *then* another set will be found to be the case — not that the first set causes the second. Thus, if the hypothesis is in fact highly probable, we can expect to

find repeated occurrences of both sets of conditions together. But the general implication says nothing about *a* causing *b*.

Second, the general implication does not assert that the consequent conditions are true. Rather, it says that *if* the antecedent conditions are true, *then* the consequent conditions are true. For example, the statement, "*If I go downtown today, then I will be robbed*" does not mean that I *will be* robbed. Even if the compound statement is true, I might not go downtown today. Thus, *if* the hypothesis is highly probable, then whether or not I will be robbed depends on whether or not I satisfy the antecedent conditions.

TYPES OF HYPOTHESES

We have suggested that the general implication is a good form for stating hypotheses. We have discussed the use of the implication, but have said nothing explicitly about the generality of the implication. In one of the previously cited examples it was said that, if an industrial work group has a certain characteristic, certain consequences follow. We did not specify what industrial work group, but it was understood that the hypothesis concerns at least *some* industrial work groups out of all possible groups. Are we now justified in asserting that the hypothesis holds for all industrial work groups? The answer to this question is ambiguous, and there are two possible courses. First, we could say that the particular work group out of all possible work groups with which we are concerned is unspecified, thus leaving the matter up in the air. Or, second, we could assume that we are asserting a universal hypothesis, i.e., that it is implicitly understood that we are talking about *all* industrial work groups that are in conflict. In this instance, if you take *any* industrial group in conflict, the consequences specified by the hypothesis should follow. In the interest of the advancement of knowledge, we lean toward the latter interpretation, for if the former interpretation is followed, no definite commitment is made on the part of the scientist, and if nothing is risked, nothing is gained. If it is found that the hypothesis is not universal in scope (that it does not apply to *all* industrial work groups), it must be further limited. This is at least a definite step forward. That this is not an idle question is made apparent by reviewing the psychological literature. Hull, for example, states a number of his empirical generalizations in this manner. His Postulate IV says: "If reinforcements follow each other at evenly distributed intervals . . . the resulting habit will increase in strength . . ." (Hull, 1952, p. 6). Without worrying here about what the specific variables in Postulate IV mean, we may observe the form of the principle. Is it clear that Hull is asserting some relationship between *all* reinforcements and *all* habits? It is by no means, but the most efficient course open to us, as we have discussed it above, would be to *assume* that he is asserting such a universal relationship.

While it is recognized that the goal of the scientist is to assert his hypotheses in as universal a fashion as possible, it is also clear that he should explicitly state the degree of generality with which he is asserting his hypotheses. With this in mind, let us investigate the possible types of hypotheses that the scientist has at his disposal.

The first type of hypothesis is what is called the *universal hypothesis*, which asserts that the relationship in question holds for all the variables that are specified, for all time, and at all places. An example of a universal hypothesis would be "For all rats, if they are rewarded for turning left, then they will turn left in a T maze." We may add that universal hypotheses in psychology typically have to be restricted in scope (see discussion, pp. 45-46).

Another type of hypothesis is the *existential hypothesis*, which asserts that the relationship stated in the hypothesis holds for at least one particular case ("existential" implies that one exists). For instance, "There is at least one rat, that if he is rewarded for turning left, then he will turn left in a T maze."³ Examples of this type of hypothesis abound. Hull's Postulate V (D), for instance, says, "At least some drive conditions tend partially to motivate into action habits which have been set up on the basis of different drive conditions" (Hull, 1952, p. 7). Again, we need not be concerned here with what the postulate actually says, but merely note that it is in the form of the existential hypothesis. (It might be better to say, "At least one drive condition . . ." to make it clear that Hull's hypothesis assumes the existential form as previously discussed.) It may be added that this form of hypothesis can be very useful in psychological work, for many times a psychologist asserts that a given phenomenon exists, regardless of how frequently it occurs. In this connection Bugelski says, "often one subject is as useful as many, if the problem involved is of the 'is-it-possible' nature. Hermann Ebbinghaus used himself as a subject and contributed greatly to our knowledge of learning. Raymond Dodge studied his own knee jerk for several years and made notable contributions to our information about reflex action. It takes only one positive case to prove something can happen" (Bugelski, 1951, pp. 115-116).

Once a scientist has confirmed an existential hypothesis, he has accomplished his immediate task. But more than establishing the existence of a phenomenon he may wish to entertain the question of the phenomenon's generality. Typically, phenomena specified in existential hypotheses are difficult to observe, and one cannot easily leap from this type of highly specialized hypothesis to an unlimited, universal one. Rather, the scientist seeks to establish the conditions under which the phenomenon does and does not occur so that he can eventually assert a universal hypothesis with necessary qualifying conditions. Let us illustrate by oversimplifying a research project

³There are other types of hypotheses that could be discussed but because they are relatively rarely used they will not be considered here. (cf. Hempel, 1945; Reichenbach, 1949).

that sought to enhance our understanding of hallucinations (McGuigan, 1966). In the problem-formulation stage the author asked some clinical psychologists about the mechanisms of this most interesting phenomenon. The initial reply was that hallucinations were "ideational in nature." Perhaps the quizzical look on the author's face communicated the difficulty he was having about how to operationally define "ideational in nature," thus stimulating a reply that was more concrete: "Well, they're cortical events." Though buried deep in the central nervous system, hallucinations so conceived at least had some sort of potential reality status. Their direct study was another problem. A more feasible approach was to consider behavioral (muscular response) manifestations of hallucinations, regardless of possible central nervous system involvement. But even behavioral aspects of hallucinations would be difficult to record, since such behavior would clearly not be overt and readily observable. Rather, any responses involved in the occurrence of hallucinations would have to be so small that one could not see them with the naked eye; they would, therefore, have to be electronically amplified to make them "visible." Furthermore since there are a number of different kinds of hallucinations, the location of these small, covert responses would probably differ according to variety of the hallucination. One could hypothesize, for instance, that covert *speech* responses might be involved in the occurrence of auditory hallucinations, that covert *ocular* responses might uniquely occur in the case of visual hallucinations, and so forth. Even leaving aside the difficulties of recording covert responses, there are questions of how to obtain patients who will admit that they hallucinate (most psychotics who evidently hallucinate refuse to so admit on the grounds that their physician might think that they are "crazy"), or how to know when a person actually experiences these private events, or how to adequately communicate with a schizophrenic who hallucinates. Perhaps by now you are getting a "feel" for the difficulty of this kind of research, and especially for testing a suitable universal hypothesis about covert responses in all patients who have all kinds of hallucinations. Consequently, rather than attempt to take a giant step that would be extremely difficult, if not impossible, the more modest approach in this research was to test an existential hypothesis. The notion was, so to speak, that auditory hallucinations are the product, at least in part, of a person covertly speaking (slightly "whispering") to himself. More precisely, the existential hypothesis was that "There is at least one paranoid schizophrenic who, if he auditorally hallucinates, then he emits covert oral responses." The research *did* confirm this hypothesis, i.e., it was found that slight speech responses coincided with the patient's report of "hearing voices." Once it was established that the phenomenon existed and that it could be recorded, the credibility of some sort of universal hypothesis increased; the question is just how a universal hypothesis should be qualified. To answer this question one would next attempt to record covert oral re-

sponses during the hallucinations of other patients. No doubt failure should sometimes be expected and the phenomenon might be observable, for instance only in paranoid schizophrenics who have auditory hallucinations and not for visual, olfactory, etc. hallucinators. Furthermore, success might occur only, say, for "new" patients, and not for chronic psychotics. But, whatever the specific conditions under which the phenomenon occurs, research should eventually lead to a universal hypothesis which includes a statement that limits its domain of application. For instance, it might say that "For all paranoid schizophrenics who will admit to hallucinations of the auditory variety and who have been institutionalized for less than a year, if they auditorally hallucinate, then they emit covert oral responses."

We can thus see how research progresses in a piecemeal, step by step fashion. Its goal is to formulate propositions of a general nature, but this is accomplished by studying one specific case after another, only gradually arriving at statements of increasing generality.

One of the reasons that we seek to establish universal statements is that the more general statement has the greater predictive power. Put the other way, a very specific statement has extremely limited predictive power. Consider the question, for example, of whether purple elephants exist. Certainly no one would care to assert that all elephants are purple, but it would be most interesting if one such phenomenon were observed; the appropriate hypothesis, therefore, is of the existential type. Should it be established that the existential hypothesis was confirmed, the delimiting of conditions might lead to the universal hypothesis that "For all elephants, if they are in a certain location, are 106 years old, and answer to the name 'Tony,' then they are purple." Though, such a highly specific hypothesis would clearly not be very useful for predicting future occurrences -- an elephant that showed up in that location at some time in the distant future would be unlikely to have the characteristics specified.

ARRIVING AT A HYPOTHESIS

It is difficult to specify the process by which we arrive at a hypothesis with any finality. Although considerable research is being conducted on this problem, it is not possible at this time to specify adequately just what phases a scientist goes through in arriving at a hypothesis. We can say that the process of arriving at a hypothesis is a creative matter that has been the object of studies in thinking, imagination, concept formation, and the like. This leads to a distinction advanced by Reichenbach (1938), who says that the manner in which the scientist actually arrives at his hypothesis falls within the *context of discovery*, and the presentation of the proof that a hypothesis is probably true is in the *context of justification*. Thus, science in general is not interested in the context of discovery (however, psychology is interested,

since this is a portion of its subject matter). Science is concerned with the context of justification, for here, instead of presenting the thought processes as they occurred in the development of the hypothesis, the scientist reconstructs his thinking logically; he sets forth a justifiable set of inferences that lead from one statement to another. The adequate expression of our thoughts in science is through the rational reconstruction of those thoughts. When the scientist publishes his hypothesis, and material related to it, he does not relate how he actually arrived at the hypothesis. He does not say that "while I was sitting in the bathtub the following hypothesis occurred to me" Rather he justifies his hypothesis. What he writes falls within the context of justification, which is the concern of the major part of this book. In this connection we might note that some critics of the study of scientific methodology (or philosophy of science) say that such an endeavor is worthless. They say that scientific discoveries are not made in a logical, step-by-step fashion in strict conformity with the scientific method; a scientist does not sit down with a problem and rationally go through the phases of the scientific method as listed in Chapter One. To this criticism we must answer that this may be true, or in fact it may not be true, but whether it is true or false is really irrelevant. What is relevant is that when the scientist communicates his findings to others, he utilizes the context of justification. Whether or not the use of the context of justification will facilitate the discovery of hypotheses is an empirical question for studies in the psychology of thinking. It is possible that the making of valuable discoveries can be facilitated by studying the scientific method, for it seems reasonable that if we learn how scientific discoveries have been made in the past, and set such procedures down in a systematic way, we may be able to make new discoveries more efficiently in the future.

Dealing further with the context of discovery, we may note that when a scientist arrives at a hypothesis he surveys a mass of data (implicitly or explicitly), abstracts certain aspects of it, sees some similarities in the abstractions, and relates the similarities in order to arrive at generalizations. For instance, the psychologist particularly observes stimulus and response events. He notes that some stimuli are similar to other stimuli and that some responses are similar to other responses. He defines as belonging to the same class those stimuli that he has noticed as being similar according to a certain characteristic, and similarly for the responses. Consider the following situation in a Skinner Box. A rat presses a lever and receives a pellet of food. At about the time the rat presses the lever, a click is sounded. After a number of associations between the click, pressing the lever, and eating the pellet, the rat will learn to press the lever when a click is sounded.

In this situation the experimenter judges that all of the separate instances of the lever-pressing response are sufficiently similar for them to be classified together. He thus forms a class of lever-pressing responses out of a number of

similar lever-pressing response instances. In like manner he forms a class of all of the stimulus instances of clicks, judging that all of the clicks are similar enough to form a general class. It can thus be seen that the psychologist uses classification to distribute a large amount of data into a smaller number of categories that can be handled efficiently. He facilitates the handling of these data by assigning symbols to the classes. He then attempts to formulate relationships between the classes. By "guessing" that a certain relationship exists between the classes, he formulates his hypothesis. For example, he might state that when a click is made, a certain response will follow. The hypothesis would be: If a click stimulus is presented a number of times to a rat in a Skinner Box, and if the click is frequently associated with pressing a lever and eating a pellet, then the rat will press the lever in response to the click on future occasions. Parenthetically, this seems to be typical of the process that the scientist goes through, more or less as an ideal. Some scientists, probably the more compulsive ones, go through each step in considerable detail, while others do so in a more haphazard fashion. But whether or not scientists go through these steps in arriving at a hypothesis, explicitly, they all seem to approximate them to some extent.

We may note the views of some others on this topic. Homer Dubs writes: "It is a well known fact that most hypotheses are derived from analogy. . . . Indeed, careful investigations will very likely show that all philosophic theories are developed analogues" (Dubs, 1930, p. 131). In support he points out that Locke's conception of simple and complex ideas was probably suggested by the theory of chemical atoms and compounds that was becoming prominent in his day. Underwood has written: "How does one learn to theorize? It is a good guess that we learn this skill in the same manner we learn anything else — by practice" (Underwood, 1949, p. 17).

Of course, some hypotheses are more difficult to formulate than others. It seems reasonable to say that the more general a hypothesis is, the more difficult it is to conceive. The important general hypotheses must await the genius to proclaim them, at which time a science usually makes a sizable spurt forward, as happened in the cases of, say, Newton and Einstein. It appears that, to formulate useful and valuable hypotheses, a scientist needs, first, sufficient experience in the area and, second, the quality of "genius." The main problem in formulating hypotheses in complex and disorderly areas is the difficulty of establishing a new "set" — the ability to create a new solution that runs counter to, or on a different plane from, the existing knowledge. This is where scientific "genius" is required.

Consider the source of hypotheses from a somewhat different position, in particular with reference to the results of a scientific inquiry. The findings can be regarded, in a sense, as a stimulus to formulate new hypotheses — while results may be used to test a hypothesis, they can also suggest additional ones. For example, if the results indicate that the hypothesis is false, they

can possibly be used to form a new hypothesis that is in accord with the results of the experiment. In this case the new hypothesis must be tested in a new experiment. Now, what happens to a hypothesis if it turns out to be false? If there is a new (potentially better) hypothesis to take its place, it can be readily discarded. But if there is no new hypothesis, then we are likely to maintain the false hypothesis, at least temporarily, for no hypothesis ever seems to be finally discarded in science unless it is replaced by a new one.

CRITERIA OF HYPOTHESES

After we have formulated our hypothesis (or better, our hypotheses), we must determine whether or not the hypothesis is a "good" one. Of course, we eventually will test our hypothesis to determine whether the data confirm or disconfirm it, and certainly, other things being equal, a confirmed hypothesis is better than a disconfirmed hypothesis, in the sense that it offers one solution to a problem and thus provides some additional knowledge about nature. But even so, some confirmed hypotheses are better than other confirmed hypotheses. We must now inquire into what we here mean by "good" and by "better". To answer this question we offer the following criteria by which to judge hypotheses. Each criterion should be read with the understanding that the hypothesis that meets it to the greatest extent is the best hypothesis, assuming that the hypothesis satisfies the other criteria equally well and that the criteria are about equal in worth. It should also be understood that these are flexible criteria. They are offered tentatively, and as the information in this important area increases, they will no doubt be modified. The hypothesis:

1. . . . must be testable. The hypothesis that is presently testable is superior to the hypothesis that is only potentially testable.
2. . . . should be in general harmony with other hypotheses in the field of investigation. While this is not essential, in general the disharmonious hypothesis has the lower degree of probability. For example, a medical doctor recently advanced the hypothesis that eye color is related to certain personality characteristics. This hypothesis is at an immediate disadvantage because it conflicts with the existing body of knowledge. We know, for instance, that hair color has never been demonstrated to be related to personality traits. There is considerable additional knowledge of this sort which, taken together, would suggest that the "eye color" hypothesis is not true — it is not in harmony with what we already know.
3. . . . should be parsimonious. If two hypotheses are advanced to answer a given problem, the more parsimonious one is to be preferred. For example, if we have evidence that a person has correctly guessed the symbol (hearts, clubs, diamonds, spades) on a number of cards significantly more

often than chance, we might advance several hypotheses to account for this phenomenon. One might be to postulate extrasensory perception (ESP) and another to say that the subject "peeked" in some manner. Clearly the latter would be more parsimonious. This principle, or ones similar to it, has been previously expressed in various forms. For instance, William of Occam advanced a rule (called *Occam's razor*) to the effect that entities should not be multiplied without necessity. A similar rule was expressed by G. W. Leibniz' principle of the identity of indiscernibles. Lloyd Morgan's canon is an application of the principle of parsimony to psychology: "In no case is an animal activity to be interpreted in terms of higher psychological processes, if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development" (Morgan, 1906, p. 59). It is apparent that these three principles have the same general purpose — they all seek, other things being equal, the most parsimonious explanation of a problem (cf. Newbury, 1954). Thus, we should not prefer a complex hypothesis if a simple one has equal explanatory power; we should not use a complex concept in a hypothesis (e.g., ESP) if a simpler one will serve as well (e.g., peeking at the cards); we should not ascribe higher capacities to organisms if the postulation of lower ones can equally well account for the behavior to be explained.

4. . . . should answer (be relevant to) the problem. It would seem unnecessary to state this criterion, except that examples can be found in the history of science where the right answer was given to the wrong problem. It is often important to make the obvious explicit.

5. . . . should have logical simplicity. By this we mean logical unity and comprehensiveness, not ease of comprehension. Thus, if one hypothesis can account for a problem by itself, and another hypothesis can also account for the problem but requires a number of supporting hypotheses or ad hoc assumptions, the former is to be preferred because of its greater logical simplicity (cf., Cohen and Nagel, 1934, pp. 212-15). (The close relationship of this criterion to that of parsimony should be noted.)

6. . . . should be expressed in a quantified form, or be susceptible to convenient quantification. The hypothesis that is more highly quantified is to be preferred. The advantages of a quantified over a nonquantified hypothesis was illustrated earlier in the example from the work of Becquerel.

7. . . . should have a large number of consequences, should be general in scope. The hypothesis that yields a large number of deductions (consequences) will explain more facts that are already established, and will make more predictions about events that are as yet unstudied or unestablished (some of which may be unexpected and novel). In general it may be said that the hypothesis that leads to the larger number of deductions will be the more fruitful hypothesis.

THE GUIDANCE FUNCTION
OF HYPOTHESES

We have already discussed the ways in which hypotheses allow us to establish "truth." It is well here to ask how an inquiry gets its direction. In nature, for instance, how do we know where to start our search for "truth"? The answer is that hypotheses direct us. An inquiry cannot proceed until there is a suggested solution to a problem in the form of some kind of hypothesis.⁴

Francis Bacon proposed that the task of the scientist is to classify the entire universe. But the number of data in the universe is, if not infinite, at least indefinitely large. To make observations in such a complex world, we must have some kind of guide. Otherwise we would have little reason for not sitting down where we are and describing a handful of pebbles or whatever else happens to be near us. We must set some priority on the kind of data that we study, and this is accomplished by the hypothesis. Hypotheses, then, serve to guide us to make observations that are pertinent to our problem; they tell us which observations are to be made and which observations are to be omitted. If, for instance, we are interested in the problem of why a person taps every third telephone pole he passes, our hypothesis would probably guide us in the direction of a better understanding of compulsions. It would take us a long time if we started out in a random direction to solve our problem and commenced, for instance, counting the number of blades of grass in a field. That this point is not universally accepted, however, is apparent when one notes that others have asserted that hypotheses are not valuable. They ask, for instance, what there is to guide us in selecting our hypotheses. To this we must answer that we know very little, but that we must be confident that more complete answers are, at least eventually, forthcoming as our research on the thinking process accumulates. Such a question is one that clearly lies within the context of discovery, and to place it within the context of justification is a serious error.

Whether or not one chooses to assume that a hypothesis guides an inquiry in a strictly exploratory type of experiment (see p. 61) is an arbitrary decision. The author is taking the position here that some guidance is offered in this type of situation, that there is some reason that data of a particular kind are gathered, and whether or not there is an explicit hypothesis, it is assumed that there is at least an implicit one, no matter how vague it might be. As Underwood says "... probably no one undertakes an investigation without some thought as to 'what will happen.' Such thoughts may not be verbalized, but that they are almost universally there no one can doubt" (Underwood, 1957, p. 208). And again, "Research in a relatively new area of investigation is seldom undertaken without some conceptual scheme in mind . . . without some preconception as to the nature of the phenomena and perhaps the processes lying behind them. These predilections are usually lightly held but they do afford the initial working hypotheses . . ." (Underwood, 1957, p. 258).

ON ACCIDENT, SERENDIPITY,
AND HYPOTHESES

The goal of understanding that part of nature which is called "the behavior of organisms" is probably as difficult a task as man has ever set himself. One of the reasons for its difficulty is the great expanse of the behavioral realm; the number of response events that we *could* conceivably study staggers the imagination. Consequently, we need to assign priorities to the kinds of behavioral phenomena that we study in any detail. Hypotheses, we have said, serve this function — they help to tell us which of an indefinitely large number of responses are more likely to justify our attention. During the conduct of an experiment in which a certain hypothesis is being tested, however, one need not be blind to other events. In fact, to reach his goal the psychologist must necessarily be alert to all manner of happenings other than that to which he is *primarily* directing his attention. For, on occasion, some chance observation that is irrelevant to, or different from, the hypothesis being tested may lead to the formulation of an even more important hypothesis. We have, you will note, mentioned several examples of the role of accidental observations in science. This is, in fact, a sufficiently important phenomenon that it justifies the use of a unique term. "Serendipity" is the word that has recently become quite popular. "Serendipity" was borrowed from Walpole's "Three Princes of Serendip" by the physiologist Cannon (1945). Walpole's story concerned a futile search for something, but the finding of many valuable things which were not sought. And so it is in science — the scientist may vainly seek "truth" by being guided by one hypothesis, but in his search accidentally observe an event that leads to a more fruitful hypothesis. An interesting case in point is related by Fisher (1964). This researcher was interested in setting off drives such as hunger and thirst by direct chemical stimulation of specific brain cells. It had been established that the injection of a salt solution into the hypothalamus of goats increased the thirst drive, thus resulting in their drinking large quantities of water. Analogously, Fisher sought to test the hypothesis that injection of the male sex hormone into a rat's brain would trigger male sexual behavior. As he tells the story of the "The Case of the Mixed-up Rat":

By one of those ironic twists that are so typical of scientific research, the behavioral change produced in my first successful subject was a completely unexpected one. Within seconds after the male hormone was injected into his brain he began to show signs of extreme restlessness. I then put in his cage a female rat that was not in the sexually receptive state. According to the script I had in mind, the brain injection of male hormone should have driven the male to make sexual advances, although normally he would not do so with a nonreceptive female. The rat, however, followed a script of his own. He grasped the female by the tail with his teeth and dragged her across the cage to a corner. She scurried away as soon as he let go, whereupon he dragged her back again.

After several such experiences the male picked her up by the loose skin on her back, carried her to the corner and dropped her there.

I was utterly perplexed and so, no doubt, was the female rat. I finally guessed that the male was carrying on a bizarre form of maternal behavior. To test this surmise I deposited some newborn rat pups and strips of paper in the middle of the cage. The male promptly used the paper to build a nest in a corner and then carried the pups to the nest. I picked up the paper and pups and scattered them around the cage; the male responded by rebuilding the nest and retrieving the young.

After about 30 minutes the rat stopped behaving like a new mother; apparently the effect of the injected hormone had worn off. Given a new injection, he immediately returned to his adopted family. With successive lapses and reinjections, his behavior became disorganized; he engaged in all the same maternal activities, but in a haphazard, meaningless order. After an overnight rest, however, a new injection the next day elicited the well-patterned motherly behavior.

The case of the mixed-up male rat was a most auspicious one. Although the rat had not followed the experimenter's script, the result of this first experiment was highly exciting. It was an encouraging indication that the control of behavior by specific neural systems in the brain could indeed be investigated by chemical means. We proceeded next to a long series of experiments to verify that the behavior in each case was actually attributable to a specific chemical implanted at a specific site in the brain rather than to some more general factor such as mechanical stimulation, general excitation of the brain cells, or changes in acidity or osmotic pressure (Fisher, 1964, pp. 2-4).

Fisher's research is another good example of serendipity and well illustrates the flexibility that is characteristic of the successful scientist. What is being urged, then, is that while one is testing a hypothesis he continue to be alert to accidental occurrences that will stimulate other research. Almost without exception, in fact, the experimenter who patiently and flexibly observes his subjects gets many hints for the development of hypotheses other than the one he is presently testing. The position has been taken, though, that scientists should not, or do not, explicitly test hypotheses (cf., Sidman, 1960). This may strike you as an extreme position, but the advocacy of it has been quite explicit. Bachrach (1965), in referring to this matter, advances his first law that "People don't usually do research the way people who write books about research say that people do research" (Bachrach, 1965, ix). More specifically, he coins the phrase "*hypothesis myopia*, a common disease among researchers holding certain preconceived ideas that might get in the way of discovery" (Bachrach, 1965, p. 22). The argument, in short, is that if one seeks to test a hypothesis, he may thereby be blinded to events other than those related to the hypothesis, and these other events are potentially very important.

The primary fault with this type of argument is that it erroneously places the blame on the hypothesis, not where the blame properly belongs. We have taken the position that hypotheses are valuable, that the research of any

experimenter is guided by some sort of hypothesis, even though that hypothesis consists only of vague notions. And regardless of the precision with which a hypothesis is stated, the possession of a hypothesis does not *per se* blind the experimenter from making observations other than those called for by his hypothesis. The hypothesis, furthermore, is not the only "set of preconceived ideas that might get in the way of scientific discovery." All manner of biases, we have seen (pp. 33-34), may operate against scientific discovery. In short the term *hypothesis myopia* should be replaced with *experimenter myopia*. To exorcise the hypothesis from scientific research is to throw the baby out with the bath. Our experimentation can, in all probability, best proceed by explicitly formulating and testing hypotheses; at the same time we should keep alert for accidental occurrences that may lead to the development of even more valuable hypotheses. It only needs to be added that overemphasis of the role of accident in scientific discovery also has its dangers. One cannot, for example, enter his laboratory with confidence that "serendipity will save the day." The hard facts of everyday experimentation are that the extremely large majority of accidental occurrences have little significance — the investigation of each odd little rise in an extended performance curve, the inquiry into every consequence of equipment malfunction on a subject's behavior, the quest for the "why" of every unscheduled response can consume all of an experimenter's energy. He must, at least most of the time, keep his eye on the doughnut and not on the hole.

Let us now review what we have said. In this chapter we defined a hypothesis, a tentative solution to a problem, as a testable statement of a potential relationship between two or more variables. Statements, we said, fall into one of three categories. The analytic statement is one that is formed in such a way that it is necessarily true, but empty. The contradictory statement is also empty, and it is necessarily false. Hypotheses are synthetic statements, for they are neither absolutely true or false — rather, they have a determinable degree of probability, and they are not empty in that they are our attempts to say something about nature. The most prominent type of hypothesis is the universal one, and it is, at least ideally, stated in the form of a general implication. Existential hypotheses, those that state that there is at least one phenomenon that has a given characteristic, are useful in science too.

The manner in which a scientist arrives at a hypothesis falls within the context of discovery, and the manner in which he reports his scientific reasoning that justifies his conclusions falls within the context of justification. We also considered how we can evaluate hypotheses and, to this end, advanced seven tentative criteria for judging hypotheses. One of the functions of hypotheses is that they give direction to the experimenter. They indicate which kinds of data he should collect, and which kinds should be ignored, though one should constantly be alert for hints from his data that might lead to other hypotheses.

With this background, we are now ready to consider, in some detail, the methods by which the experimenter subjects his hypothesis to empirical test. After he has formulated his problem and the proposed solution to it, the experimenter turns his attention to the plan for conducting his experiment.

THE EXPERIMENTAL PLAN

THE EVIDENCE REPORT

We have noted that a scientific inquiry starts with a problem. The problem must be *solvable* and may be stated in the form of a question (Chapter Two). The inquiry then proceeds with the formulation of one or several hypotheses as possible solutions to the problem (Chapter Three). In the next phase of the scientific investigation the hypothesis (or the hypotheses if there are more than one) is tested to determine whether it is probably true or false. This amounts to conducting a study in which certain empirical results that relate to the hypothesis are obtained. The results of the study are then summarized in the form of an *evidence report*. For our immediate purposes we shall simply consider that an evidence report is a summary statement of the results of an investigation, i.e., it is a sentence which concisely states what was found in the inquiry. Once the evidence report has been formed it is related to the hypothesis. By comparing the hypothesis (the prediction of how the results of the experiment will turn out) with the evidence report (the statement of how the

results *did* turn out), it is possible to determine whether the hypothesis is probably true or false. We now need to inquire into the various methods in psychology of obtaining data that may be used to arrive at an evidence report. This amounts to an inquiry into the various methods of scientific investigation in psychology.

METHODS OF OBTAINING THE EVIDENCE REPORT

The methods to be discussed have in common the fact that they allow the investigator the opportunity to collect data from which, regardless of the specific method used to obtain them, an evidence report can be formulated.

NONEXPERIMENTAL METHODS

Since this is a book on experimental methodology we should not consider nonexperimental methods to any great extent. But to enhance our perspective of experimental methods, it is valuable to briefly consider the nonexperimental ones, too. The manner of classifying the nonexperimental methods is somewhat arbitrary and varies with the classifying authority. There are two general types that can be contrasted with the experimental method.

The first general type of method to consider is the *clinical method* (sometimes called the "case-history method" or the "life-history method"). The psychologist uses this method in an attempt to help the individual solve his problems, be they emotional, vocational, or whatever. In the most general form of the clinical method, the psychologist attempts to help by collecting information about the person from every possible source. He is interested in all aspects of a person's life, from birth to the present time. Some of the techniques for collecting this information might be an intensive interview, perusing records, administering psychological tests, questioning other people about the individual, studying written works of the person, or obtaining biographical questionnaires. Then, on the basis of such information, the psychologist tries to determine the factors that led to the development of the person's problem. This leads to the formulation of an informal hypothesis as to the cause of the person's problem; and the collection of further data will help him to determine whether the hypothesis is probably true or false. Once the problem and the factors that led to its development are laid bare for the person, the psychologist will try to help his subject achieve a better adjustment to the circumstances. It should be noted that the clinical method is generally used in an applied, as against a basic sense, since its usual aim is to solve a practical problem, not to advance science. (See p. 327). However observation of behavior through this method is occasionally a source of

more general hypotheses that can be taken out of the clinical setting into the laboratory for stringent testing.

The second general method is that of *systematic observation* or the *field study* method. In using this method the investigator goes into the "field" to collect his data. He takes an event as it occurs naturally and studies it, with no effort to produce or to control the event, as in experimentation. The observation of children at free play would be one example of the use of this method. The purpose there might be to determine what kinds of skills children of a certain age possess. A number of types of play apparatus would be made available for the children, and their behavior would be observed and recorded as they played. Another example of the use of the method of systematic observation might be a study of panic. We do not ordinarily produce panic in groups of people for psychological study. Rather, psychologists usually wait until a panic occurs naturally, and then set out to study it. An example of how social psychologists studied a panic was after the Orson Welles' radio dramatization of H. G. Wells' *War of the Worlds*. Psychologists were interested in why the panic occurred and thus went into the field and interviewed people who participated in it (Cantril, 1940).

The data obtained by the method of systematic observation can be used for testing hypotheses through the construction of evidence reports, but it should be obvious that rather important limitations exist. Such limitations will be briefly discussed after a consideration of the experimental method.

THE EXPERIMENT

In the early stages of the development of a science, the nonexperimental methods tend to be more prominent. In some sciences, sociology, for example, there is little hope that anything but nonexperimental methods can be generally used. This is primarily because sociology is largely concerned with the effect of the prevailing culture and social institutions on behavior, and it is difficult to manipulate these two factors as independent variables in an experiment. In those fields that are ultimately susceptible to experimentation, however, a change in methodology eventually occurs as knowledge accumulates. Hull (1943, p. 1) points out that as scientific investigations become more and more searching, the "spontaneous" happenings in nature are not adequate to permit the necessary observations. This leads to the setting up of special conditions to bring about the desired events under circumstances favorable for scientific observations, and experiments thus originate. Thus, the experimenter takes an active part in producing the event. Some advantages of this approach are well expressed by Woodworth and Schlosburg:

1. The experimenter can make the event occur when he wishes. So he can be fully *prepared* for accurate observation.

2. He can *repeat* his observation under the same conditions for verification; and he can describe his conditions and enable other experimenters to duplicate them and make an independent check of his results.

3. He can *vary the conditions* systematically and note the variation in results . . . (1955, p. 2).

Since psychology is concerned with the behavior of organisms, in using nonexperimental methods the psychologist must wait until the behavior in which he is interested occurs naturally. The researcher does not have control over the variable that he wishes to study; he can only observe it in its natural state. The one characteristic that all the nonexperimental methods have in common is that the variables that are being evaluated are not purposefully manipulated by the researcher.¹ It follows that when a theory is tested through the use of the experimental method, the conclusion is more highly regarded than if it is tested by a nonexperimental method. Put another way, the evidence report obtained through experimentation is more reliable than that obtained through the use of a nonexperimental method. This is true largely because the interpretation of the results is clearer in an experiment. Ambiguous interpretation of results can occur in nonexperimental methods primarily because of a lack of control over extraneous variables. It is difficult, or frequently impossible, in a systematic observation study to be sure that the findings, with respect to the dependent variable, are due to the independent variable, for they may result from some uncontrolled extraneous variable that happened to be present in the study. In nonexperimental methods it is also usually more difficult to define the variables studied than where they are actually produced, as in an experiment.

All of this does not mean, however, that the experimental method is a perfect method for answering questions. Certainly it can lead to errors and in the hands of poor experimenters the errors are sometimes great. Relatively speaking, however, the experimental method is preferred *where it can be appropriately used*. But if it cannot be used, then we must do the next best thing and use the method of systematic observation. Thus when it is not possible to produce the events that we wish to study, as in the example of panic, we must rely on nonexperimental methods. But we must not forget that when events

¹To emphasize this definition of an experiment, suppose that a researcher is interested in the way that learning speed changes with age. He might have a group of 20-year-old people and a group of 60-year-old people. Both groups would learn the same task. While at first glance this might appear to be an "experiment," we would not so classify it because the researcher did not *purposely manipulate* his "independent variable" (age of the subjects). Rather, he *selected* his subjects to fit certain age requirements. It is apparent that "age of subjects" is not a variable over which a researcher has control. He cannot say to one subject, "You will be 20 years old," and to another subject: "You will be 60 years old." Hence, this investigation typifies the use of the method of systematic observation. Put another way, as we shall see, in an experiment, subjects are randomly assigned to the experimental and control groups. But in the method of systematic observation, they are not.

are *selected* for study, rather than being *produced* and *controlled*, caution must be exercised in accepting the result.

One frequent criticism of the experimental method is that when an event is brought into the laboratory for study (as it usually is) the nature of the event is thereby changed. For one thing, the event does not naturally occur in isolation, as it is made to occur (relatively so) in the laboratory, for in natural life there are always many other variables that influence it. Criticism of experiments on such grounds is unjustifiable, since we really want to know what the event is like when it is uninfluenced by other events. It is then possible to transfer the event back to its natural situation at which time we know more about how it is produced. The fact that an event may appear to be different in the natural situation, as compared to the laboratory, simply means that it is being influenced by other variables, *which in their turn* need to be brought into the laboratory for investigation. Once all the relevant variables that exist under natural conditions have been studied in isolation in the laboratory, and it has been determined in what way they all influence the dependent variable and each other, then a thorough understanding of the natural event will have been accomplished. Without such an analysis of events in the laboratory, it is likely that we would never be able adequately to understand them.

Now, while we have made some strong statements in favor of the laboratory analysis of events, we must recognize the possibility that they actually may be changed or even "destroyed." This occurs when the experimenter has simply not been successful in transferring the event into the laboratory. He has produced a truly different event than that which he wished to study. For instance, it may be that if a subject knows that he is in an experiment, he may actually act differently than he would in "real life." The best that can be said for this situation is that usually events cannot be studied adequately unless they are brought into this laboratory, and at least in the laboratory suitable controls can be introduced so that even if the event is distorted by observation, the event is studied with this effect held constant.

Skinner (1953, p. 435) makes a similar point when he says, in essence, that certain characteristics of behavior are too complex to understand through casual observation of everyday life — to find the relevant variables that determine a certain kind of behavior would require considerable time, and even then it may not be possible to find them through such a common sense approach. But with adequate recording devices in the laboratory, and under controlled conditions, we can determine the variables that are responsible for an event. These findings can be utilized to good advantage in further study of the complex world at large. To illustrate his point, Skinner suggests that casual observation has led the layman to conclude that a particularly undesirable type of behavior may be eliminated by punishing a person. Punishment gives quick results; a particular type of behavior seems to dis-

appear immediately if its perpetrator is punished. However, actual laboratory studies have indicated that punishment is rather ineffective in eliminating a response. While it may affect a temporary suppression, it does not seem to permanently eliminate behavior, even when it is applied over a long period of time. A more effective technique for eliminating a response is the process of extinction, but the discovery of this finding could not have been made from the observation of everyday life. It required laboratory investigation.

TYPES OF EXPERIMENTS

In your reading you may run across a number of terms that refer to different types of experiments; to prevent possible confusion some of them will be discussed here. First, note well that the *general experimental method* remains the same, regardless of the type of problem to which it is applied, so that, strictly speaking, we are not talking about different types of experiments, but different types of problems and purposes for which they are used. To clarify this matter, contrast *exploratory* with *confirmatory* experiments. The type the experimenter uses depends on the state of the knowledge relevant to the problem with which he is dealing. If there is little knowledge about a given problem, the experimenter performs an exploratory experiment. Lacking much knowledge about the problem he is usually not in a position to formulate a possible solution — generally, he cannot postulate an explicit hypothesis that might guide him to predict such and such a happening. He is simply curious, collects some data, but doesn't really have any basis on which to guess how the experiment will turn out. Although he does not have an explicit hypothesis, he evidently has an "informal" hypothesis, at least to the extent that he has decided to investigate the effect of one specific variable on another, rather than any variable to a host of other variables. But his hypothesis is not sufficiently advanced to say what kind of effect the one variable will have on the other, or even to say that there *will* be an effect. It can be seen that the exploratory experiment is performed in the earlier stages of the investigation of a problem area. As he gathers data relevant to the problem, the experimenter becomes increasingly capable of formulating hypotheses of a more clear-cut nature. He is able to predict, on the basis of a hypothesis, that such and such an event should occur. At this stage of knowledge development he performs the confirmatory experiment, i.e., he starts with an explicit hypothesis that he wishes to test. On the basis of that hypothesis he is able to predict an outcome of his experiment; he sets up the experiment to determine whether the outcome is, indeed, that predicted by his hypothesis. Put another way, in the exploratory experiment the scientist

is interested primarily in finding new independent variables that affect a given dependent variable, while in the confirmatory experiment he is interested in confirming that a given variable is influential. In the confirmatory experiment he may also want to determine the extent and precise way in which one variable influences the other, or more generally, to determine the functional (quantitative) relationship between the two variables. Underwood (1949, pp. 11-14) used two descriptive terms to refer to the problems which the two types of experiments are used to solve. The exploratory experiment refers to his "I-wonder-what-would-happen-if-I-did-this" type of problem, while the confirmatory experiment is analogous to his "I'll-bet-this-would-happen-if-I-did-this" type of problem.

Regardless of the state of knowledge in a given problem area, the immediate purpose of an experiment is to arrive at an evidence report. If the experiment is exploratory, the evidence report can be used as the basis for formulating a specific, precise hypothesis. In a confirmatory experiment the evidence report is used to determine whether the hypothesis is probably true or false. In the latter case, if the hypothesis is not in accord with the evidence report, it might be modified to better fit the data and then tested in a new experiment. If the hypothesis is supported by the evidence report, then its probability of being true is increased. In addition to providing such a general understanding, the distinction between these two types of experiments has direct implications for the types of experimental designs that are employed, one type of design being more efficient for the exploratory experiment, while another is more efficient for the confirmatory experiment. These designs will be discussed later.

Sometimes you may run across the term "crucial experiment" (*experimentum crucis*). This term is used to describe an experiment that purports to test one or several "counter-hypotheses" simultaneously. For instance, it may be possible to design the experiment in such a manner that if the results come out one way, one hypothesis can be said to be confirmed and a second hypothesis disconfirmed, and if the results point another way, the first hypothesis is said to be disconfirmed and the second hypothesis is confirmed. Ideally, a crucial experiment is one whose results support one theory, and disconfirm all possible alternative theories. However, we can seldom if ever be sure that we can state at any given time all of the possible alternative hypotheses. Accordingly, perhaps we can never have a crucial experiment, but only approximations of one.

The term *pilot study* or *pilot experiment* has nothing to do with the behavior of aircraft operators, as one student thought, but refers to a preliminary experiment, one conducted prior to the major experiment. It is used, usually with only a small number of subjects, to suggest what specific values should be assigned to the variables being studied, to try out certain procedures to see how well they work, and more generally to find out what mistakes might be

made in conducting the actual experiment so that the experimenter can be ready for them. It is a dress rehearsal of the main performance.

PLANNING AN EXPERIMENT

Given a problem to be solved and a hypothesis, which is a tentative solution to that problem, we must design an experiment that will determine whether that hypothesis is probably true or false. In designing an experiment the researcher uses his ingenuity to obtain data that are relevant to the hypothesis. This involves such problems of experimental technique as: What apparatus will best allow manipulation and observation of the phenomenon of interest? What extraneous variables may contaminate the phenomenon of primary interest and are therefore in need of control? Which events should be observed and which should be ignored? How can the results of the experiment best be observed and recorded? By considering these and similar problems an attempt is made to rule out the possibility of collecting irrelevant evidence. For instance, if the antecedent conditions of the hypothesis are not satisfied, the evidence report will be irrelevant to the hypothesis, and further progress in the inquiry is prohibited. Put another way, the hypothesis says that *if* such and such is the case (the antecedent conditions of the hypothesis), *then* such and such should happen (the consequent conditions of the hypothesis). The hypothesis amounts to a contract that the experimenter has signed — he has agreed to make sure that the antecedent conditions are fulfilled. If he fails to fulfill his agreement, then whatever results he collects will have nothing to do with his hypothesis; they will be irrelevant and thus cannot be used to test the truth of the hypothesis. This points up the importance of adequately planning the experiment. For if the experiment is improperly designed, then either no inferences can be made from the results, or it may only be possible to make inferences to answer questions that the experimenter has not asked. That this is not an idle warning is indicated by the frequency with which the right answer is given to the wrong question, particularly by neophyte experimenters. And if the only result of an experiment is that the experimenter learns that he should not make these same errors in the future, this is very expensive education indeed.

It is a good idea for the experimenter to draft a rather thorough plan of the experiment before he conducts it. Once the complete plan of the experiment is set down on paper it is desirable to obtain as much criticism of it as possible. The experimenter often overlooks many important points, or looks at them with a wrong, preconceived set, and the critical review of others may bring potential errors to the surface. No scientist is beyond criticism, and it is far better for him to accept criticism before an experiment is conducted than to

make errors that might invalidate the experiment. We shall now suggest a series of steps that the experimenter can follow in the planning of an experiment.

OUTLINE FOR AN EXPERIMENTAL PLAN

1. *Label the Experiment.* The title should be clearly specified, as well as the time and location of the experiment. As time passes and the experimenter accumulates a number of experiments in the same problem area, he can always refer to this information without much chance of confusing one experiment with another.

2. *Survey of the Literature.* All of the previous work that is relevant to the experiment should be studied. This is a particularly important phase in the experimental plan for a number of reasons. First, it helps in the formulation of the problem. The experimenter's vague notion of what problem he wants to investigate is frequently made more concrete by consulting other studies. Or, the experimenter thus may be led to modify his original problem in such a way that the experiment becomes more valuable. Another reason for this survey of pertinent knowledge is that it tells the experimenter whether or not the experiment even needs to be conducted. If essentially the same experiment has previously been conducted by somebody else, there is certainly no point in repeating the operation, unless it is specifically designed to confirm previous findings. Other studies in the same area also provide numerous suggestions about extraneous variables that need to be controlled and hints on how to control them.

The importance of the literature survey cannot be overemphasized. The experimenters who tend to slight it usually pay a penalty in the form of errors in the design or some other complication. The knowledge in psychology is growing all the time, making it more difficult for one person to comprehend the findings in any given problem area. Therefore, this step requires particularly close attention. Also, since reference to relevant studies should be made in the write up of the experiment, this might just as well be done before the experiment is conducted thus combining two steps in one.

We are very fortunate in psychology to have the *Psychological Abstracts*, which makes any such survey relatively easy.² Every student of psychology should attempt to develop a facility in using the *Abstracts*.

3. *Statement of the Problem.* The experiment is being conducted because

²The *Psychological Abstracts* is a professional journal published monthly by the American Psychological Association. It summarizes the large majority of psychological research and classifies it according to topics (and authors) so that it is fairly easy to determine what has previously been done on any given problem. The December issue of each year summarizes all the research for that year.

there is a lack of knowledge about something. The statement of the problem expresses this lack of knowledge. Although the problem can be developed in some detail, through a series of logical steps, the actual statement of the experimental question should be concise. It should be stated succinctly and unambiguously in a single sentence, preferably as a question. The statement of the problem as a question implies that it can be answered unambiguously in either a positive or negative manner. If the question cannot be so answered, *in general* we can say that the experiment should not be conducted. Every worthwhile experiment involves a gamble. If the problem cannot be definitely answered either positively or negatively, the experimenter has not risked anything and therefore cannot hope to gain new knowledge.

4. *Statement of the Hypothesis.*—The variables specified in the statement of the problem are explicitly stated in the hypothesis as a sentence. Natural languages (e.g., English) are usually employed for this purpose, but other languages (e.g., mathematical or logical ones) can also be used, and, in fact, are preferable. The “if . . . then . . .” relationship was suggested as the basic form for stating hypotheses.

5. *Definition of Variables.* The independent and dependent variables have been specified in the statement of the problem and of the hypothesis. They must now be defined in such a manner that they are clear and unambiguous — the variables must be *operationally defined*. The importance of this phase has been previously emphasized (pp. 27–28). To repeat here, if no such operation is possible for all the experimental variables, we must conclude that the hypothesis is untestable.

6. *Apparatus.* Every experiment involves two things: (1) an independent variable must be manipulated; and (2) the resulting value of the dependent variable must be recorded. Perhaps the most frequently occurring type of independent variable in psychology is the presentation of certain values of a stimulus; and in every experiment a response is recorded. Both of these functions may be performed manually by the experimenter. However, it is frequently desirable, and in fact sometimes necessary to resort to mechanical or electrical assistance. We may even make the bold statement that the more an experimenter can rely on apparatus in his experiment, the better off he will be. There are naturally some exceptions to this general statement, but if the apparatus is adequate for the job, those exceptions will be few. Thus, there are two general functions of apparatus in psychological experimentation: (1) to facilitate the administration of the experimental treatment; and (2) to aid in recording the resulting behavior. Let us consider how these two functions might be advantageously accomplished in an experiment.

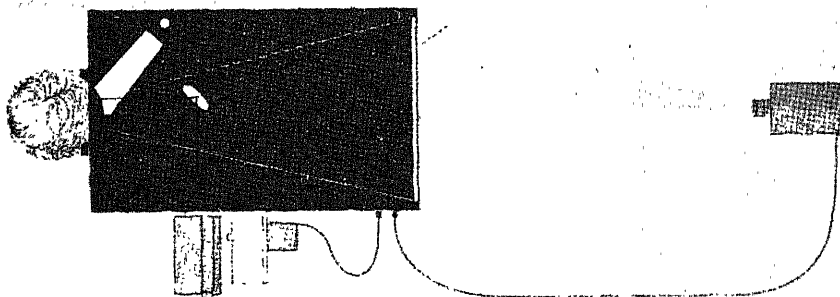
Expressions from literature and from everyday language have long alluded to the emotional significance of the size of the pupils of the eyes, as, for instance, “His eyes were like saucers” or “His eyes were pinpoints of hate.” Hess (1965) has recently conducted a series of studies in which the topic has

been systematically investigated. His general procedure has been to present a variety of stimuli and to relate these stimuli to resulting changes in pupillary size. His subjects peer into a box, looking at a screen on which is projected a



FIGURE 4.1.

(for a): Subject in pupil-response studies peers into a box, looking at a rear-projection screen on which slides are flashed from the projector at right. A motor-driven camera mounted on the box makes a continuous record of pupil size at the rate of two frames a second.



(for b): Pupil-Response apparatus is simple. The lamp and the camera film work in the infrared. A timer advances the projector every 10 seconds, flashing a control slide and a stimulus slide alternately. The mirror is below eye level so that view of screen is clear.

Figure 4.1, a and b, from Eckhard H. Hess, "Attitude and Pupil Size." Copyright © 1965 by *Scientific American, Inc.* All rights reserved. Photograph (4.1a): Courtesy of Saul Mednick.

stimulus picture, as shown in Fig. 4.1. A mirror reflects the image of the subject's eye into a motion picture camera. First a control slide is present for ten seconds; this slide is matched in overall brightness with an experimental slide so that the subject's eyes are adapted to the appropriate light intensity. Then the experimental slide is presented for ten seconds, and this sequence of alternating control and experimental slides is continued for ten or twelve times a sitting. To quantify the dependent variable, the movie film is projected on a screen and the size of the pupil is measured either with a ruler or electronically by means of a photo cell. The results have been most intriguing: In general, interesting or pleasant pictures lead to dilation of the pupils, while unpleasant or distasteful stimuli lead to pupillary constriction. Thus the presentation of a female pinup produces greater enlargement of the pupils of men than it does of women; but women showed a greater enlargement than did men to a male pinup, or to a picture of a mother and a baby. Distasteful pictures, such as of sharks or of crippled or cross-eyed children, resulted in a decrease of pupillary size. Other studies showed that when an arithmetic problem is presented, the size of the pupil increases to a maximum at the point at which the solution is reached and returns to its base level as soon as the answer is reported.

Let us now return to the main point we wish to illustrate through these studies. We may note that apparatus was used to present the experimental treatment — a device for presenting a stimulus for whatever length of exposure the experimenter desired. In this case pictures and control slides were projected for ten seconds. Furthermore, there was an automatic timing device with a driving motor so that experimental and control slides were automatically alternated every ten seconds. The second function of apparatus in experimentation was fulfilled by the movie camera. It operated at a rate of two frames per second, so that pupillary size was photographed regularly (Fig. 4.2). The use of a ruler or photo cell then allowed quantification of the dependent variable.

The types of apparatus used in behavioral experimentation are so numerous that we cannot attempt a systematic coverage of them here.³ We shall, however, briefly try to illustrate further the value of apparatus, as well as to refer to certain cautions that should be observed.

Frequently a certain stimulus, such as a light, must be presented to a subject at very short intervals. It would be difficult for an experimenter to time the intervals manually and thus make the light come on at precisely the desired moments. In addition to the considerable error that would be in-

³An excellent general source is the book by Grings (1954); for a more detailed presentation of electronic instrumentation see Brown & Saucer (1958). A general discussion of instrumentation with specialized articles by experts in various fields is offered by Sidowski (1966). Cornsweet (1963) concentrates on the design of electrical circuits, while Yanof (1965) and Venables and Martin (1967) take up the area of biomedical electronics (psychophysiology, etc.).

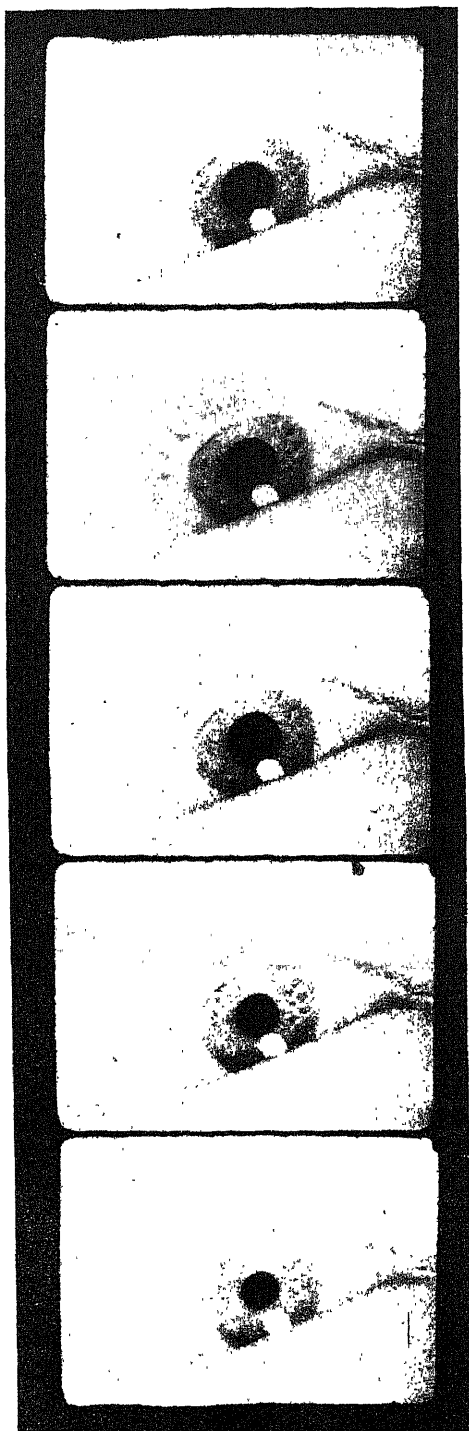


FIGURE 4.2.

Pupil size varies with the interest value of a visual stimulus. The five frames show the eye of a male subject during the first 2 1/2 seconds after a photograph of a pinup girl appeared on the screen. His pupil diameter increased in size 30 per cent.

From Eckhard H. Hess, "Attitude and Pupil Size." Copyright © 1965 by *Scientific American, Inc.* All rights reserved.

volved, the work required of the experimenter would be highly undesirable, and in fact might distract him from other important aspects of the experiment. This type of stimulus presentation could be handled very easily by placing a metronome in the circuit to break and complete the circuit at the proper times. As you can see, one of the main advantages of apparatus is that it reduces the "personal equation." Suppose that a reaction-time experiment is being conducted. For example, one might wish to present a series of words, one at a time, to a subject and measure the time it takes him to respond with the first word that comes to mind. If the experimenter were forced to measure the subject's reaction time by starting a stop watch when the word is read and stopping it when the subject responds, considerable error would result. The reaction time of the experimenter in starting and stopping the stop watch enters into the "subject's reaction time;" this is especially bad because the experimenter's reaction time would vary and hence a constant value could not be subtracted as a correction factor. A better approach would be to use a voice key that is connected in circuit with a timing device. In this apparatus, when the word is read to the subject, the timer automatically starts, and when the subject responds, it automatically stops. The experimenter may then record the reaction time and say the next word. While the "reaction time" of the apparatus is still involved, it is at least constant and of smaller magnitude, as compared to the experimenter's reaction time. A similar example of the valuable use of apparatus would be in timing a rat as he runs a maze. Several types of apparatus that automatically record the rat's latency and running time have been developed for this problem.

We have been emphasizing the *desirability* of apparatus in experimentation, but there are a large number of problems that simply cannot be studied without the use of apparatus. In some experiments apparatus is not only valuable, it is downright essential. Probably the clearest illustration would be the recording of response measures of a psychophysiological nature. If one wishes to study the effect of some independent variable on electroencephalograms ("brainwaves"), an electroencephalograph is required. Similarly, for measuring pulse rate a sphygmograph is necessary, for recording galvanic skin responses a psychogalvanometer needs to be used, and so forth.

In spite of all the advantages of apparatus, there are a number of possible disadvantages. We have assumed above that, to be of value, the apparatus is suitable for the job required of it. This is not always the case, for sometimes apparatus is not accurate, not adequately calibrated, etc. Furthermore, sometimes apparatus may interfere with the event being studied. Problems of this nature are discussed in greater detail in the chapter on control (Chapter Six). But there is one potential disadvantage that should be mentioned here, a disadvantage that can be illustrated by analogy with the "Law of the Hammer." This law, which is a genuinely universal one, states simply that

if you give a little boy a hammer, he will find many things that need pounding. And so it is, at least to some extent, with the scientist — he somewhat too frequently finds many things that *need* to be recorded with the particular apparatus that he has in his laboratory. In this way the availability of the apparatus can determine the problem that is researched when, clearly, it should be the other way around. While no doubt a boy does find some items that really require pounding, many hammered items are satisfactory prior to a juvenile onslaught. The more fruitful approach is for the scientist to formulate his problem and then consider his equipment requirements. Otherwise he might not only needlessly expend energy on unfruitful projects, but might also be somewhat blinded to more valuable research areas.

7. *Control of Variables.* In this phase of planning the experiment the scientist must consider all of the variables that might contaminate the experiment; he should attempt to evaluate any and all variables which might affect his dependent variable. He may decide that some of these extraneous variables might act in such a way that they will invalidate the experiment, or at least leave the conclusion of the experiment open to question. Such variables need to be controlled. To elaborate on what we mean by “control” and the techniques for achieving it will demand much attention; since it is of extreme importance in experimentation, it will be the subject of a later discussion (Chapter Six). For now, let us simply note a point that will be expanded later — that we must make sure no extraneous variable will differentially affect our groups, i.e., that no such variable will affect one group differently than it will affect another group.

In the course of examining variables that may influence our dependent variable, we will have to make certain decisions. We will decide, preferably on the basis of information afforded by previous research, that a certain variable is not likely to be influential, in which case it may be reasonable to ignore it.

Other variables, however, might be considered to be relevant, but there are difficulties in controlling them. Perhaps it is reasonable to assume that such variables will exert equal effects, at random, on all conditions. If this assumption seems tenable, the experimenter might choose to proceed. But if such an assumption is rather tenuous, then the problem may well be of such serious proportions that it would be wiser to abandon the experiment.

8. *Selection of a Design.* So far we have discussed only the two-groups design, i.e., where the results for an experimental group are compared to those for a control group. We shall consider a number of other designs later, from among which the experimenter may choose the one most appropriate to his problem. For example, it may be more advantageous to use several groups, instead of just two, in which case a *multi-groups* design would be adopted. Another type of design, which in many cases is the most efficient and which is being used

more and more in psychology, is the *factorial* design. For now, however, you should restrict your consideration to the two-groups design, at the same time realizing that a number of alternatives are possible.

9. *Selection and Assignment of Subjects to Groups.* The experimenter conducts an experiment because he wants to conclude something about behavior. To do this, of course, he must select certain subjects to study. But from what group should he select his subjects? This is an important question because he will want to generalize his findings from the subjects that he is studying to the larger group of subjects from which they were chosen (see step 14, page 76). The larger group of subjects is the *population* (or “universe”) under study; those who participate in his experiment constitute his *sample*. More generally, by “population” we mean the total number of possible items of a class that might be studied — it is the entire group of items from which a sample has been taken. Thus, we may note that a population need not refer only to people, but to any type of organism: amoebae, rats, jellyfish. Furthermore, one may note that the definition is worded in such a manner that it can and does refer to inanimate objects. For example, we may have a “population” of types of therapy (directive, nondirective, etc.), of learning tasks (hitting a baseball, learning a maze, etc.), and so forth. Or an experimenter may be interested in sampling a population of stimulus conditions (high, medium, low intensity of a light), or a population of experiments (three separate experiments that test the same hypotheses, etc.). In engineering, a person may be interested in studying a population of the bridges of the world, and an industrial psychologist may be concerned with a population of whiskey products, and so on.

In designing an experiment one should specify with great precision the population (or populations) he is studying. For the moment let us merely concern ourselves with subject populations, and leave other populations until later (see Chapter Fourteen). In specifying a population we must note those of its characteristics that are particularly relevant to its definition, e.g., if we are concerned with a population of people, we might wish to specify the age, sex, education, socioeconomic status and race. If we are working with animals, we might wish to specify the species, sex, age, strain, experience, habitation procedures, and feeding schedules. Unfortunately, one can observe by reading articles in professional journals that experimenters rarely define the populations they are studying with any great precision.

Given a well-defined population that we wish to study, then, we are faced with the problem of how to actually study it. If it is a small population, it may be possible to observe each individual. And adequately studying an entire population is far preferable to studying a sample of it. The author once conducted some consumer research studies in which he was supposed to ob-

tain a sample of 18 people in a small town in the High Sierras. After considerable difficulty he was able to locate the "town" and after further difficulty he was eventually able to find an eighteenth person. In this study, the entire population of the town was exhausted (as was the author), and more reliable results were obtained than if a smaller sample of the town were selected. As it turned out, however, not a single person planned on purchasing TV sets, dishwashers, or similar electrical appliances during the next year, largely because the town did not have electricity.

In any event, the population to be studied is seldom so small that it can be exhausted by the researcher. More likely, the population is so large, sometimes infinitely large, that it cannot be studied in its entirety, and the researcher must resort to studying a sample. One of the reasons that a population may be infinitely large (or perhaps finite, but indefinitely large), is that the experimenter may wish to generalize not only to people now living, but to people who are as yet unborn.

Where the population is too large to be studied in its entirety, the experimenter must select a number of subjects and study them. One technique that an experimenter may use in selecting a sample is that of *randomization*. In random selection of a sample of subjects from a population each member of the population has an equal chance of being chosen. For instance, if we wish to draw a random sample from a college of 600 students, we might write the names of all the students on separate pieces of paper. We would then place the 600 slips of paper in a (large) hat, mix thoroughly, and, without looking, draw our sample. If our sample is to consist of 60 students, we would select 60 pieces of paper. Of course there are simpler techniques to achieve a random sample. For instance, we might take a list of all 600 students and select every tenth one to form our sample. To select the first name, we would randomly select one of the first ten subjects, and then count successive tens from there.

Once the experimenter randomly selects a sample, he then assumes that his sample is typical of the entire population — that he has drawn a *representative* sample. Drawing samples at random is usually sufficient to assure that his sample is representative, but the researcher may check on this if he wishes.⁴ For one, if he has values available for the population, then he can compare his sample with the population values. If he is studying the population of people in the United States, he has readily available a large amount of census infor-

⁴A somewhat infrequent definition of a representative sample is one that has been randomly drawn from a population. Following this definition, of course, it would be foolish to check the representativeness of the sample. We are, however, following the more common procedure of assuming that a sample that is not representative (is very atypical) of a population can be randomly selected. For example let us define a population of people that consists of 90 blonds and 10 redheads. Now if we draw a random sample of 10, and find that they are all redheads, we would say that the sample is not representative of the population. See Lindquist (1953) for an excellent discussion of this topic.

mation about education levels, age, sex, etc.⁵ He can compute certain of these statistics for his sample and compare these figures with those for the general population. If the values are close, he can assume that his sample is representative. The assumption that he is making, in this case, is that if his sample is similar to the population in a number of known characteristics, it is also similar with respect to characteristics for which no data are as yet available. This could be a dangerous assumption, but it is certainly better than if there was no check on representativeness.

Once the population has been specified, a sample drawn from it, and the type of design determined, it is necessary to divide the sample into the number of groups to be used. The subjects must be assigned to groups by some random procedure. By using randomization we would assure ourselves that each subject has an equal opportunity to be assigned to each group. Some procedure such as coin flipping can be used for this purpose. For example, suppose that we have a sample of twenty subjects, and that we have two groups. We might take the first subject and flip a coin; if it is "heads," the subject will be placed in group one, and if it is "tails," he will be placed in group two. We would then do likewise for the second subject, and so on until we have ten subjects in one group. The remaining subjects would then be assigned to the other group.

We now have two groups of subjects who have been assigned at random. The next thing to do is to determine which group is to be the experimental group and which is to be the control group. This decision should also be determined in a random manner, such as by flipping a coin. We might make a rule that if a "head" comes up, group one is the experimental group and group two the control group; but if our coin flipping yields a "tail," group one would be the control group, group two the experimental group.

By now you should have acquired a feel for the importance of randomization in experimental research — the random selection of a sample of subjects from a population, the random assignment of subjects to groups, and the random determination of which of the two groups will be the experimental group and which will be the control group. It is by the process of randomization that we attempt to eliminate biases (errors) in our experiment. When we want to make statements about our population of subjects, we generally study a sample that is representative of that population. If our sample is not representative, then what is true of our sample may not be true of our population, and we might make an error in generalizing the results obtained from our sample to the population. Random assignment of our sample to two groups is important because we want to start our experiment with groups that are essentially equal. If we do not randomly assign subjects to two

⁵Of course, some census statistics are also obtained through sampling techniques and such statistics do not guarantee that the population has those precise values. We are simply assuming here that such interval statistics are *probably* true — a safe assumption.

groups, we may well end with two groups that are unequal in some important respect. If we assign subjects to groups in a nonrandom manner, perhaps just looking at each subject and saying "I'll put you in the control group," we may have one group being more intelligent than the other; consciously or unconsciously, we may have selected the more intelligent subjects for the experimental group.

Having thus emphasized the importance of these procedures, we must hasten to add that the use of randomization does not guarantee that our sample is representative of the population from whence it came, or that the groups in an experiment are equal before the administration of the experimental treatment. For, by an unfortunate quirk of fate, randomization may produce two unequal groups, e.g., one group may, in fact, turn out to be significantly more intelligent than the other. However, randomization is typically the best procedure that we can use, and we can be sure that, at least in the long run, its use is justified. For any given sample, or in any given experiment, randomization may well result in "errors," but here, as everywhere else in life, we must play the probabilities. If a very unlikely event occurs (e.g., if the procedure of randomization leads to two unequal groups) we will end up with an erroneous conclusion. Eventually, however, due to the self-checking nature of science, the error will be discovered.

We may conclude discussion of this step of the experimental plan by noting that the number of groups to be used in an experiment is determined by the number of independent variables and the number of values of them that we have selected for study, as well as by the nature of the variables to be controlled. For instance, if we have a single independent variable that we decide to vary in two ways, we would have two groups. More than likely these two groups would be called experimental and control groups. If we select three values of the independent variable for study, then obviously we would need to assign our sample of subjects to three groups. It might be added that usually an equal number of subjects is assigned to each group. Thus if we have 80 subjects in our sample, and if we vary the independent variable in four ways, we would have four groups in the experiment, probably 20 subjects in each group. It is not necessary, however, to have the same number of subjects in each group, and the experimenter may determine the size of each group in accordance with criteria that we shall take up later.

10. *Experimental Procedure.* The procedure for conducting the data collection phase of the experiment should be set down in great detail. The experimenter should carefully plan how the subjects will be treated, how the stimuli will be administered, how the response will be observed, and recorded. He should specify the instructions to the subjects (if humans are used) and formulate a statement concerning the administration of the independent variable and the recording of the dependent variable. It is very useful to make an outline of each point to be covered in the actual data collection

phase. The experimenter might start his outline right from his greeting to the subject and carry through step by step to when he says "goodbye." It is also advisable to try out a few subjects to see how the procedure works. More often than not such "dress rehearsals" will suggest new points to be covered and modifications of procedures already set down. And if more elaborate checking of the procedure is desired, a pilot study might be conducted (see p. 62).

11. *Statistical Treatment of the Data.* The data of the experiment are usually subjected to statistical analysis. As psychology has progressed this phase of experimentation has become increasingly important. We have witnessed the development of some very powerful statistical techniques. In some manner the reliability of the results of the experiment should be evaluated. Suppose, for instance, a two-groups design is used and the sample mean for the experimental group is found to be 14.0, the sample mean for the control group 12.1. In this case one might conclude that the population mean for the experimental group is higher than for the control group. On the basis of this limited amount of information, however, such a conclusion is not justified, since it has not been determined that the sample difference is a reliable (significant) difference. The observed difference may not be a "real" difference. If the difference is not significant, the next time the experiment is conducted the outcome may be reversed. Thus, as we noted in Chapter One, we need to use a statistical technique to determine whether the difference between the mean scores of the two groups is significant. The statistical analysis will tell you, in effect, the odds that the difference between the groups might have occurred by chance. If the probability that this difference could have occurred by random fluctuations is small, then we may conclude that the difference is significant, that the experimental group is reliably superior to the control group.

The main point here is that the data are evaluated by a statistical test, a number of which are available. Some tests are appropriate to one kind of data or experimental design and some are not. The use of such statistical tests frequently requires that certain assumptions about the experimental design and the kind of data collected must be met. In taking such matters into consideration it is advisable to plan the complete procedure for statistical analysis prior to conducting the experiment. Sometimes experimenters do not do this, and find that there are serious problems in the statistical analysis which could have been prevented by a little more planning and insight. Lack of rigor in the use of statistics can invalidate the experiment. And the statistics used must be appropriate to the design selected.⁶

12. *Forming the Evidence Report.* We have said that the evidence report is a summary statement of the findings of the experiment, but we can now add

⁶A "non-statistical," though rigorous, approach will be presented later (Chapter Eleven).

that it should tell us something more than this. It should also tell us that the antecedent conditions of the hypothesis held (were actually present) in the experiment. More completely, then, the evidence report is a statement that asserts that the antecedent conditions of the hypothesis obtained, and that the consequent conditions specified by the hypothesis were found either to occur or not to occur. If the consequent conditions were found to occur, we may refer to the evidence report as positive, and if they were not found to occur, the evidence report is negative. To illustrate, let us return to a hypothesis previously considered (p. 41): "If a teacher praises a student for good reading performance, then the student's reading growth will increase." Now, let us design an experiment to test this hypothesis. The subjects in an experimental group shall be praised each time they exhibit good reading performance. No praise is given to the members of the control group when they show similar performance. Let us assume that the experimental group exhibits a significantly greater increase in reading growth than does the control group. Referring to the hypothesis, we may note that the antecedent condition was satisfied, and that the consequent condition was found to be the case. We may thus formulate our evidence report: "Students were praised by a teacher when they exhibited good reading performance, and they exhibited an increase in reading growth (as compared to the control group)." In short, the evidence report is a sentence that asserts that the antecedent conditions held *and* that the consequent conditions either did or did not hold — it is of the form "*a* and *b*," where *a* stands for the antecedent conditions and *b* for the consequent conditions of the hypothesis.

13. *Making Inferences from the Evidence Report to the Hypothesis.* In this phase the evidence report is related to the hypothesis for the purpose of determining whether the hypothesis is probably true or false. To do this we must make an inference from the evidence report to the hypothesis. Essentially the inference is the following: If the evidence report is positive, the hypothesis is confirmed (the evidence report and the hypothesis coincide — what was predicted to happen by the hypothesis actually happened, as stated by the evidence report). If, however, the evidence report is negative, we may conclude that the hypothesis is disconfirmed.

14. *Generalization of the Findings.* The extent to which the results of the experiment can be generalized depend on the extent to which the populations with which the experiment is concerned have been specified and the extent to which those populations have been represented in the experiment by random sampling. Considering only subject populations again, let us say that the experimenter specified his population as all of (and only) the students at Ivy College. If he has randomly drawn a sample of subjects from that population, his experimental results may be generalized to that population; he may assert that what was true for this sample is probably true for the whole population. Of course, if the population was not adequately defined, or the

sample was not randomly drawn from it, no such generalization can be made, and the results would apply only to the sample studied.

A SUMMARY AND PREVIEW

We have now covered the major steps of experimentation, and you should now have a good idea of the individual steps and how they fall into a logical pattern. Our first effort to present the whole picture was in Chapter One. In Chapters Two and Three and in this section we have attempted to enlarge on some of the steps. Thus in Chapter Two we considered the nature of the problem, and in Chapter Three we discussed the hypothesis. These two initial phases of planning the experiment were summarized as steps 3 and 4. Next we said that the variables specified by the hypothesis should be operationally defined (step 5). The use of apparatus for presenting stimuli and for recording responses was discussed as step 6. The important topic of control was briefly considered (step 7), but will be enlarged on later (Chapter Six). Following this we pointed out that several designs are possible in addition to the two-groups design that we have largely concentrated on (step 8). The ways in which several different experimental designs may be used is the subject of Chapters Five, Eight, Nine, Ten, and Eleven. Next we took up the selection of subjects and assignment of them to groups (step 9). Step 10 consisted of a brief discussion of experimental procedure, and in step 11 we offered a preview of the techniques of statistical analysis that will be more thoroughly covered in connection with the different experimental designs. The formation of the evidence report, and the way in which it is used to test the hypothesis were taken up in steps 12 and 13. A separate chapter will be devoted to the latter topic (Chapter Twelve). Finally, we briefly considered the problem of generalization (step 14), a topic that will be more elaborately considered in Chapters Thirteen and Fourteen. Of course, as we continue through the book, each of the above points will continue to appear in a variety of places, even though separate chapters may not be devoted to them. As a summary of this section, as well as to facilitate your planning of experiments, we offer the following check list.

1. Label the experiment.
2. Summarize previous research.
3. State your problem.
4. State your hypothesis.
5. Define your variables.
6. Specify your apparatus.
7. State the extraneous variables that need to be controlled and the ways in which you will control them.

8. Select the design most appropriate for your problem.
9. Indicate the manner of selecting your subjects, the way in which they will be assigned to groups, and the number to be in each group.
10. List the steps of your experimental procedure.
11. Specify the type of statistical analysis to be used.
12. State the possible evidence reports. Will the results tell you something about your hypothesis no matter how they come out?
13. Are you clear about the inferences that can be made from the evidence report to the hypothesis?
14. To what extent will you be able to generalize your findings?

Now that we have presented the steps of experimentation in an orderly and logical fashion, let us conclude this section with a tempering comment. Bachrach (1965) puts it something like this: If something can go wrong, it will. This point is included now because, after faithfully following the prescriptions here offered, students typically experience difficulties in the conduct of an experiment that they summarize by the phrase "everything's a mess." By this they apparently mean such things as: the equipment stopped working in the middle of an experimental session, some subjects were uncooperative, a fatal error in control of variables was detected after the data were collected, and so forth.

Now clearly such difficulties may present themselves during the conduct of an experiment, but they are just as clearly not the sole possession of students;⁷ the sophisticated researcher also experiences his share of such grief, though he has learned to be more agile and is better able to recover when troubles appear. Adjustments nearly always are made by the professional researcher, such as replacing the data for one subject for whom procedural errors occurred by running another subject under the same condition. The list of steps presented here, in short, is meant to be a help in carefully planning your experiment. And the anticipation of potential problems will help to reduce the number of experimental errors that occur. We might only add that experienced psychologists themselves profit to the extent to which they formulate and adhere to a precise experimental plan.

CONDUCTING AN EXPERIMENT — AN EXAMPLE

One of the values of the conduct of an experiment early in your course in experimental psychology is that it affords you the opportunity to make

⁷You might also be consoled by Bachrach's second law when you find yourself spending what seems like an immoderate amount of time on your research. It states simply that: "*Things take more time than they do.*" (Bachrach, 1965, ix).

errors that you can learn to avoid in later research. For this reason it is important for students to commence work on a problem early in the course, regardless of the simplicity of the experiment or of whether or not it will contribute new knowledge. Too many students feel that their first experiment has to be an important one. Certainly, we want to encourage the conduct of important research, but the best way to reach the point where this is possible is to practice. The following example is a realistic one at this stage of your training; the fact that it is a simple, straightforward one allows us to better illustrate the points that we have previously covered.

The problem that one class set themselves concerned the effect of knowledge of results on performance: They wanted to know whether informing a person of how well he performs a task will facilitate his learning of that task. The title of this experiment was "The Effect of Knowledge of Results on Performance," and the students conducted a rather thorough literature survey on that topic. The problem was then stated: "What is the effect of knowledge of results on performance?" The hypothesis was: "If knowledge of results is furnished to a person, then that person's performance will be facilitated." Note that the statement of the problem and the hypothesis has *implicitly* determined the variables; they next need to be made *explicit*. The task that the subjects performed was the drawing, while blindfolded, of 5 inch lines. The independent variable concerns the amount of knowledge of results furnished the subjects, and it was varied from zero to some large amount. A large amount of knowledge of results was operationally defined as telling the subject whether his line was "too long," "too short," or "right." "Too long," in turn, was defined as any line $5\frac{1}{4}$ inches or longer, "too short" as any line $4\frac{3}{4}$ inches or shorter, and "right" as any line between $5\frac{1}{4}$ inches and $4\frac{3}{4}$ inches. A zero amount of knowledge of results was defined as furnishing the subject no information about the length of his line. The dependent variable was the actual length of the lines that he drew. Each subject was required to draw 50 lines, his proficiency being determined by his total performance (the sum total of his deviations from 5 inch lines) on all 50 trials.

The apparatus consisted of a drawing board on which was affixed ruled paper, a blindfold, and a pencil. The paper was easily movable for each trial, and ruled in such a manner that the experimenter could tell immediately within which of the three intervals (long, short, or right) the subject's lines fell.

Two values of the independent variable were selected for study, a positive amount and a zero amount. Two groups were thus required, an experimental and a control. The experimental group received knowledge of results, whereas the control group did not. The subject population was defined as all of the students in the college. From a list of the student body,

60 subjects were randomly selected for study.⁸ The 60 subjects were then randomly divided into two groups (see p. 73 for the precise manner in which this assignment may be carried out). It was then randomly determined that one of the groups was the experimental group, and the other was the control group.

Next it was determined which extraneous variables might influence the dependent variable and therefore needed to be controlled. Our general principle concerning control is that both groups should be treated alike in all respects, except as far as the independent variable is concerned (in this case different amounts of knowledge of results were administered). Hence, essentially the same instructions were read to both groups; a constant "experimental attitude" was maintained in the presence of both groups. The experimental plan specified that the experimenter should not frown at some of the subjects and be gay or jovial with others. Incidental cues were eliminated insofar as possible, e.g., the experiment might have been invalidated had the experimenter held his breath when the subjects reached the 5 inch mark. Not only would this have furnished some knowledge of results to an alert control subject, but it would have increased the amount of knowledge of results for the experimental subjects.

Is the amount of time between trials an important variable? Previous research suggested that it was. In general, the longer the time between trials, the better the performance. This variable was therefore controlled by holding it constant for all subjects. It was specified that each subject should wait ten seconds after each response before his hand was returned to the starting point for the next trial. What other extraneous variables might be considered? Perhaps the time of day at which the subjects are run is important; a person might perform better in the morning than in the afternoon or evening. If the experimental group were run in the morning and the control group in the afternoon, then no clear-cut conclusion about the effectiveness of knowledge of results could be drawn from the data. One control for this time variable might be to run all subjects between 2 P.M. and 4 P.M. But even this might produce differences, since subjects might perform better at 2 P.M. than at 3 P.M. Furthermore, it was not possible to run all the subjects within this one hour on the same day, so the experiment had to be conducted over a period of two weeks. Now does it make a difference whether subjects are run on the first day or the last day of the two weeks? It may be that examinations are being given concurrent with the first part of the experiment,

⁸It was correctly assumed that all sixty subjects would cooperate. The fact that this assumption is not always justified leads to the widespread practice among experimenters of using students in introductory psychology classes. Such students are quite accessible to psychologists and usually "volunteer" readily. This method of selecting subjects, of course, does not result in a random sample, and thus leads to the question of whether the sample is representative of the population (all the students in the college).

causing the subjects to be nervous. Then again it may be that people who are tested on Monday perform differently than people tested on Friday.

The problem of how to control the time variable was rather complex. The following procedure was chosen (see Chapter Six for an elaboration): It was specified that all subjects would be run between 2 P.M. and 4 P.M. When the first subject reported to the laboratory, a coin was flipped to determine whether he was an experimental or a control subject. If it turned out that he was a control subject, the next subject was assigned to the experimental group. When the third subject reported, it was similarly determined which group he was assigned to and the fourth subject was placed in the other group. The rest of the subjects were similarly assigned to the groups for as many days as the experiment was conducted. By using this procedure it was rather safely assumed that whatever the effects of time differences on the subjects' performance, they were balanced — that they affected both groups equally. This is so because, in the long run, we can assume that an equal number of subjects from both groups participated during any given time interval of the day, and on any particular day of the experiment.

Another control problem concerns the individual characteristics of the experimenter, a topic that we shall explore in considerable detail later (Chapter 14). In this case all of the students in the experimental psychology class ran subjects for this experiment. Should it have been the case that one student ran more subjects than another, or that the students did not run an equal number of experimental and control subjects, experimenter characteristics might have differentially affected the dependent variable measures of the two groups. The experimenter variable was, thus, adequately controlled. The illustrations given should be sufficient to illustrate the control problems involved. Think of some additional variables that the class considered, e.g., do various distracting influences exist such as noise from radiators and people talking? These could be controlled to some extent, but not completely. In the case of those that could not be reasonably controlled, it was assumed that they affected both groups equally — that they "randomized out." For instance, there is no reason to think that the various distracting influences should affect one group to a greater extent than the other. After surveying the various extraneous variables, it was concluded that this assumption was justifiable: There were no variables that could not either be controlled or whose effect, if any, would differentially affect the dependent variable scores of the two groups.

The next step considered was the experimental procedure. The plan for this phase proceeded as follows: "After the subject enters the laboratory room and is greeted, he is seated at a table and given the following instructions: 'I want you to draw some straight lines that are five inches long, while you are blindfolded. You are to draw them horizontally like this (experimenter demonstrates by drawing a horizontal line in the air). When you have com-

pleted your line, leave your pencil at the point where you stopped. I shall return your hand to the starting point. Also keep your arm and hand off the table while drawing your line. You are to have only the point of the pencil touching the paper. Are there any questions?" The experimenter will answer any questions by repeating pertinent parts of the instructions. When the subject is ready, the experimenter blindfolds him ("now I am going to blindfold you"), uncovers the apparatus, and places the pencil in the subject's hand. The subject's hand is guided to the starting point and he is instructed: 'Ready? Go.' The experimental subjects are given the appropriate knowledge of results (as previously specified) immediately after their pencils stop. No information is given to the control subjects. After the subject completes a trial, he waits ten seconds, after which his hand is returned to the starting point. He is told: 'Now draw another line five inches long. Ready? Go.' This same procedure is followed until the subject has drawn 50 lines. The experimenter must move the paper before each trial so that the subject's next response can be recorded. The subject's blindfold is then removed, he is thanked for his cooperation, and cautioned to discuss the experiment with no one."

Following this, the students collected their data. It was reassuring, though hardly startling, to find that knowledge of results did, in fact, facilitate performance.

Illustration of the final steps of the planning and conduct of this experiment (statistical treatment of the data, forming the evidence report, confronting the hypothesis with the evidence report, and generalization of the findings) can best be offered when these topics are later emphasized.

WRITING UP AN EXPERIMENT

After the experimenter has collected his data, subjected them to statistical analysis, and reached his conclusions, he writes up the experiment. The point of view taken here is that the same general format for writing up experiments should be used regardless of whether the report is to be submitted for publication in a scientific journal or whether it is a study conducted by a beginning class in experimental psychology. This helps to maximize the transfer of learning from a course in experimental psychology to the actual conduct of experiments as professional psychologists. The following is an outline that can be used for writing up the experiment. There are also offered a number of suggestions that should help to eliminate certain errors that students frequently make, and several other suggestions that should lead them to a closer approximation to scientific writing.

The main principle to follow in writing up an experiment is that the report must include every relevant aspect of the experiment; someone else should be able to repeat the experiment solely on the basis of the report. If this is

impossible, the report is inadequate. On the other hand, the experimenter should not become excessively involved in details. Those aspects of an experiment which the experimenter judges to be irrelevant should not be included in his report. In general, then, the report should include every important aspect of the experiment, but should also be as concise as possible, for scientific writing is economical writing.

The writer should also strive for clarity of expression. If an idea can be expressed simply and clearly, it should not be expressed complexly and ambiguously; "big" words or "high flown" phrases should be avoided wherever possible.

We shall adhere to certain standard conventions. The conventions and a number of additional matters about writing up an experiment may be found in the *Publication Manual* of the American Psychological Association. The close relationship between the write-up and the outline of the experimental plan should be noted. Frequent reference should be made to that outline in the following discussion for much of the write-up has already been accomplished there.

1. *Title.* The title should be short but indicative of the exact topic of the experiment. This does not mean that every topic included in the report should be specified in the title. The title needs to be unique — it should distinguish the experiment from all other investigations. Introductory phrases such as "A Study of . . ." or "An Investigation of . . ." should be avoided, since they are generally understood.

2. *Name and Institutional Connection.* On the title page the author's name should be centered below the title, and the next line should state his institutional connection. In the case of multiple authorship where all authors are from the same institution, the affiliation should be listed last (and only once). In no case should the department within the institution be specified. Frequently an entire class conducts an experiment, in which case, strictly speaking, they are multiple authors. Since the main purpose of such class experiments is to provide practice for the individual student, however, it is suggested that only the name of the student writing up the experiment be used as the author, rather than listing the entire class including the professor.

3. *Introduction.* It was noted previously that the write-up of the literature survey could serve as the basis for the introductory section of the report. We said that the problem should be developed logically, citing the most relevant studies. A summary statement of the problem should then be made, preferably as a question. Let us emphasize that the results of the literature survey should lead quite smoothly into the statement of the problem. For instance, if the experiment concerns the effects of alcohol on performance of a cancellation task (e.g., striking out all letter E's in a series of letters), you might cite the results of previous experiments that show detrimental effects of alcohol on various kinds of performance. At this point you might indicate

that there is no previous work on the effects of alcohol on the cancellation task and that the purpose of your experiment was to extend the previous findings to that task. Accordingly the problem is, "Does the consumption of alcohol detrimentally affect performance on a cancellation task?" The steps leading up to the statement of the hypothesis should also be logically presented, but it too should be stated in one sentence, preferably in the "If . . . then . . ." form. It is not customary to label the introductory section; rather, it should simply start as the first part of the article.

4. *Method.* The main function of this section of the report is to tell your reader precisely how the experiment was conducted. Put another way, this section serves to specify the method of gathering data that are relevant to the hypothesis and that will serve to test the hypothesis. It is here that the main decisions need to be made as to which matters of procedure are relevant and which are irrelevant. If the author has specified every detail that is necessary for someone else to repeat the experiment, but no more, he has been successful. To illustrate, let us assume that a "rat" study has been conducted. The author would want to tell the reader that, say, a T maze was used, and then go on to specify the precise dimensions of the maze, the colors used to paint it, the type of doors, and the kind of covering. He would presumably not want to relate that the maze was constructed of pine, or that the wood used was one inch thick, for it is highly unlikely that these variables would influence the subject's performance. That is, it would be a strange phenomenon indeed if one could show that rats performed differently in a T maze depending on whether the maze (well painted) was constructed of pine, redwood, or walnut, or whether the walls were $3/4$ inch or one inch thick.

The outline used in presenting the method is not rigid and may be modified for each experiment. In general, however, the following information should be found, and usually in the following order:

a. *Subjects.* The population should be specified in detail, as well as the method of drawing the sample studied. If any subjects from the sample had to be "discarded" (students didn't show up for their appointments, they couldn't perform the experimental task, rats died, etc.), this information should be included, for the sample may not be random because of these factors.

b. *Apparatus.* All relevant aspects of the apparatus should be included. Where a standard type of apparatus is used (e.g., a "Hull-type Memory Drum"), only its name need be stated. Otherwise, the apparatus has to be described in sufficient detail for another experimenter to obtain or construct it. It is good practice for the student to include a diagram of the apparatus in the write-up, although in professional journals this is only done where the apparatus is complex and novel.

c. *Design.* The type of design used should be included in a section after the apparatus has been described. The method of assigning subjects to

groups, and the labels attached to the groups, are both indicated (e.g., Group *E* may be the experimental group and Group *C* the control group, etc.). The variables contained in the hypothesis need to be (operationally) defined; it is also desirable (at least for your practice) to indicate which are the independent and dependent variables. The techniques of exercising experimental control may be included here. For example, if there was a particularly knotty variable that needed to be controlled, the techniques used for this purpose may be discussed.⁹

d. *Procedure.* The procedure for conducting the data collection phase of the experiment should be set down in detail. You must include or summarize instructions to the subjects (if they are human), the maintenance schedule and the way in which the subjects were "adapted" to the experiment (if they are infrahuman animals), how the independent variable was administered, and how the dependent variable was recorded.

5. *Results.* The data relevant to the test of the hypothesis are presented here. These data are summarized as a precise sentence (the evidence report). If the data are in accord with the hypothesis, then it may be concluded that the hypothesis is confirmed. If they are not of the nature predicted by the hypothesis, then the hypothesis is disconfirmed.¹⁰

It is quite important to present a *summary* of the data under "results." This can almost always be done by using a table, but frequently figures can

⁹The inclusion of the definition of variables, experimental control, etc. in this section is arbitrary. You may well include separate sections for these matters, consider them under "Procedure," or arrange them otherwise.

¹⁰We have not yet reached certain important matters that need to be included in this section. This advance information is summarized below for future reference. You need not worry about what this information means if it is unfamiliar, for it will become clear later. The advantages of including an outline of all relevant information in one place seem to outweigh the disadvantage of prematurely presenting this information. You should state the null hypothesis as it applies to your experiment and also the significance level that you have adopted. Then include the results of the statistical test and indicate the appropriate probability level. For example: "The null hypothesis was: There is no difference between the means of the experimental and control groups on the dependent variable. The *t* test yielded a value of 2.20. With 16 degrees of freedom, this value is significant beyond the 5 per cent level." (An alternative way of saying this would be: "The resulting *t* was 2.20 ($P < .05$)," in which case we may assume that the degrees of freedom are obvious from the number of subjects used. However, for practice, it is a good idea for you to specify the number of degrees of freedom used. You may then continue: "It is therefore possible to reject the null hypothesis. Since this finding is in accord with the empirical hypothesis, we may conclude that that hypothesis is confirmed." Of course, if the empirical hypothesis predicted that the null hypothesis would be rejected, but it was not, then it may be concluded that the hypothesis was not confirmed.

Having made a point about the null hypothesis, let us immediately direct your attention to the fact that the null hypothesis is not mentioned in journal articles. Rather, what we have above made explicit is, for professional experimenters, implicitly understood. Perhaps your understanding of the null hypothesis can be enhanced should your write-up specifically include the above mentioned steps, and once this process is clear to you, it can be dropped from later reports.

also be used to advantage. Whether or not tables and/or figures are used depends on the type of data and the ingenuity and motivation of the writer. Since students frequently are confused about the definitions of *table* and *figure*, as well as about their respective formats, we shall consider them in detail. Tables and figures are used for *summarizing* the data. They are not used for presenting all the data (so-called "raw data," a term that implies that the data have not been statistically treated). Nor are the steps in computing the statistical tests (the actual calculations) included under "results." In student write-ups, however, it has been found advisable to include the raw data and the steps in the computation of the statistical tests in a special appendix. The advantage of using such an appendix is that the instructor can correct any errors made in this part of the operation.

A table in the results section consists of numbers that summarize the main findings of the experiment. It should present these numbers systematically, precisely, and economically. A figure, on the other hand, is a graph, chart, photograph, or like material. It is particularly appropriate for certain kinds of data; for instance, to show the progress of learning. Information should, however, be presented only once, i.e., the same data should not be presented in a table and a figure or in the written text.

In constructing a table, one should first determine what is to be shown — the main points that should be made apparent from the table. Then should be considered the question of what is the most economical way of making these points in a meaningful fashion. Since the main point of the experiment is to determine if certain relationships exist between certain variables, the table should show whether or not these relationships exist. Of course, it is possible to present more than one table, and tables may be used for purposes other than presenting data. For example, it is frequently possible to make the over-all design of the experiment more apparent by presenting the separate steps in tabular form (this use of tables is particularly recommended for students as it helps to "pull the experiment together" for them). To illustrate the format of a table, let us consider an experiment in which the effects of human environment on the cognitive ability of rhesus monkeys were studied (Singh, 1966). This experimenter compared the problem solving behavior of a group of *urban* monkeys with a group of *forest* monkeys; the environmental difference between the two groups was that the former had frequent and intimate interactions with human beings, while the latter (having lived in the jungles) did not. Both groups were administered a variety of tests, among which was one on visual pattern discrimination and another on object discrimination. The apparatus used was such that the animal was presented with two or more stimuli, and under one was a raisin. If he reached out of his cage and displaced the correct stimulus, he received a raisin. For the visual pattern discrimination test, the animal had to respond until he made 45 out of 50 correct responses in one day. When he had thus

successfully learned to visually discriminate one pattern, he was presented with another, then another, until he learned to discriminate eight patterns.

To summarize his data, Singh counted the number of trials that each animal took before he reached the criterion that showed that he had learned the discrimination. Then he determined the median number of trials for each group. The results are presented in Table 4.1. For example, it can be seen that the median number of trials required by the Urban group to discriminate the first visual pattern was 338.0; the median number for the Forest group was 491.5. Similar comparisons can be made for each of the remaining patterns. Note that the previously stated requirements of a good table are clearly satisfied in this example. Also observe the precise format used, for students have a habit of ignoring the details of the standardized conventions illustrated here.

Table 4.1. *Illustration of a Good Format for a Table.*

Table 1. *Median Trials to Criterion on Successive Visual Pattern Discriminations.*

GROUP	PATTERN							
	I	II	III	IV	V	VI	VII	VIII
Urban	338.0	149.5	24.0	0.0	165.0	26.5	0.0	0.0
Forest	491.5	261.5	34.0	40.0	259.0	102.0	44.5	13.0

By studying Table 4.1 and other tables throughout the book, as well as those in journal articles in your library, you will begin to acquire a "feel" for efficiently and systematically presenting your data. In some cases you will want to include the numbers of subjects in your groups; most frequently you will probably use means for your dependent variable, rather than medians; you will often want to include some measure of variability, such as standard deviations; and so forth.

The same general principles stated for the construction of tables holds for figures. In particular, a figure is used primarily to illustrate a relationship between the independent and dependent variables of an experiment. It is conventional to use the vertical axis (sometimes erroneously, as you can see in a dictionary, called the "ordinate") for plotting the dependent variable scores; the horizontal axis (which is *not* synonymous with "abscissa") is typically labelled "Time" or "Number of Trials." The scores for each group of subjects may then be plotted and compared. As an example of a good use of a figure, let us consider how Singh presented his data on the object discrimination problem. He determined a total number of responses made by each group of monkeys for the first 48 problems that they solved. The number of correct responses out of the total number was then counted, and the

percentage of correct responses was computed. In this way, then, it was ascertained that about 58 per cent of the total number of responses made by the Urban group to the first 48 problems were correct. During the solution of the next 48 problems, the percentage of correct responses rose; the value plotted in Fig. 4.3 is approximately 69 per cent. By studying Fig. 4.3 we can

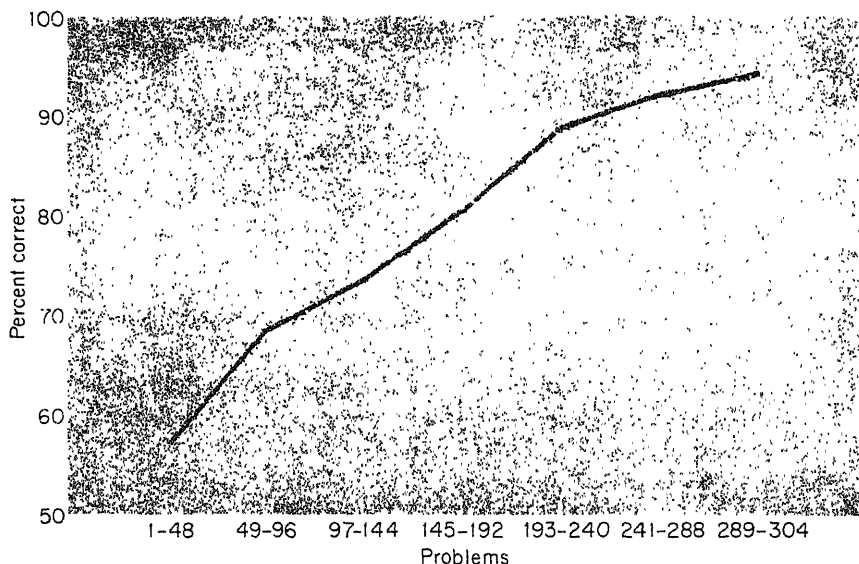


FIGURE 4.3.

Object Discrimination Learning Curves for the Urban and Forest Groups (After Singh, 1966).

see that the percentage of correct responses increases as the number of problems that the animals had solved increases; by the time the animals had solved over 300 problems their proficiency in solving new problems was considerably better than when they were naive. A comparison of the learning curve for the urban with that of the forest group shows that they are quite similar. In fact, Fig. 4.3 is illustrative of the general conclusion of this study, viz., "... the results ... do not indicate any effect of human environment on the cognitive ability of the rhesus monkeys ..." (Singh, 1966, p. 283).

Incidentally, one question that often comes up when comparing performance curves of two groups concerns what would happen if more trials had been given. For instance, suppose that the curve for an experimental group rises in a normal manner, but that the curve for the control group is retarded. Further, suppose that the experimenter gave his subjects 100 trials, and that at about trial number 90 the control group's curve markedly increases, though by trial number 100 it is still lower than that of the experimental group. What would have happened had more trials been given? Would

the two curves eventually come together? If we run the experiment again and give the subjects a larger number of test trials, we probably would find out. (Would 20 more trials be enough?) The question of what would happen to the relative position of the curves if the subjects had been run on more trials seems to be a perennial one in experimentation. One might accept this as a lesson in planning his experiment: If he is going to be concerned about this question, take it into consideration before the data are collected.

The order of presenting tables and figures is important. A table of means, or a figure in which means are plotted, is used to demonstrate your major experimental effects. A table presenting your statistical analysis (such as your analysis of variance, as discussed in Chapter Nine) is for the purpose of stating whether or not your means are significantly different. Hence, the statistical analysis should come *after* the means. It is also important to emphasize that the source of the numbers that are presented in your tables and figures be *precisely* identified and explained. Often a reader must spend considerable time puzzling over the question of just what the numbers mean — although they may seem clear to the author, his write-up may have missed a step. For instance, rather than saying that “the mean number of errors” is plotted in Figure 2, one could be more precise and say that “The mean number of errors per ten-trial block” is plotted. Or, in another case, rather than merely referring to “the number of bar-presses,” one should say “The median number of bar-presses during the 15 minute extinction period.” This information may be presented in the text, in the table heading, or in the figure caption.¹¹

As far as format is concerned, each table and figure goes on a separate page and is included at the end of the paper. There is a separate page which includes all figure captions, but table headings go at the top of the table. The author indicates where tables and figures should be located in the text as follows:

Insert Table 1 about here

The above information should be sufficient to get you started, but you are strongly advised to continue your study of techniques for constructing figures and tables. To do this, you can consult any of a number of elementary statistics books that are available in your library (e.g., Guilford, 1965; McNemar, 1962). But more important, you should concentrate on figures and tables as they are presented in psychological journals.

¹¹Thanks to Professor David A. Grant, who has read many manuscripts for the *Journal of Experimental Psychology*, for these and several other points.

6. *Discussion.* The main functions of this section are to interpret the results of the experiment and to relate these results to other studies. The interpretation involves essentially an attempt to explain the results. Perhaps some existing theory can be brought to bear to help understand the findings. If the hypothesis was derived from general theory, then the confirmation of the hypothesis serves to strengthen that theory, and the findings in turn are explained by that hypothesis in conjunction with the larger theory. If the findings are contrary to the hypothesis, then some new explanation is required; to account for the results there may be advanced new hypotheses that run counter to the hypothesis tested. Or it may be that the hypothesis tested can be modified in such a way to make it consistent with the results. In this case a "patched up" hypothesis is advanced for future test.

In relating the results to other studies, the literature survey may again be brought to bear. By considering the present results along with previous ones new insights may be obtained. The results of this particular experiment may provide the one missing piece that allows the solution of the puzzle.

Hypotheses may also be advanced about any unusual deviation in the results. For instance, one may wonder why there was a sudden "bump" in a learning curve on trial 7. Is this a reliable "bump"? If it is, why did it occur? In short, what additional problems were uncovered that require further investigation?

There may be certain limitations in the experiment. If so, this is the place to discuss them, e.g., what variables might have turned out to have been inadequately controlled? (If, however, you did not control a crucial extraneous variable, you probably wouldn't attempt to publish your report.) How would one modify the experiment if it were to be repeated?

You may also consider the extent to which the results may be generalized. To what populations may you safely generalize? To what extent are the generalizations limited by uncontrolled variables, etc.?

A rather strange characteristic of some experimenters seems to be that they feel "guilty" or "embarrassed" when they have obtained negative results.¹² Whatever the reason, it is not appropriate to include long "alibis" for negative results. It is reasonable, however, to briefly speculate about why they were obtained.

7. *References.* References to pertinent studies throughout the text should be made by citing the author's (or authors') last name, the year of publication, and enclosing these in parentheses. If the name of the author occurs in the text, cite only the year of publication in the parentheses. The references should then be listed alphabetically at the end of the paper. The form and order of items for journal references is as follows: last name, initials, title of the study,

¹²"Negative results" occur when the hypothesis makes a particular prediction, but the results are contrary to that prediction. The term may also be used to indicate that the null hypothesis (p. 105) was not rejected; usually these two definitions amount to the same thing.

the (non-abbreviated) name of the journal, year of publication of the study, volume number, and pages. For example, if two references like the following are used they might be referred to in the write-up as "According to Lewis (1953), learning theory has already been shown to have applicability to human behavior. Examples of this position are numerous (e.g., Calvin, Perkins, & Hoffman, 1956)." Then, in the "reference" section of the write-up, they would be *precisely* listed as:

References

Calvin, A. D., Perkins, M. J., & Hoffman, F. K. The effect of non-differential reward and non-reward on discriminative learning in children. *Child Development*, 1956, 27, 439-446.

Lewis, D. J. Rats and Men. *American Journal of Sociology*, 1953, 59, 131-135.

8. *Abstract.* The abstract should be the first page of your article, but it is listed last here because it is preferable to write it after you have completed the foregoing sections. This section should be typed on a separate sheet of paper; it includes the title and author (as does the first page of the text) and generally summarizes the article. Its function is to quickly give the reader the essence of your research. More particularly, the abstract should be about 100 to 120 words in length and "... should contain statements of (a) the problem, (b) the method, (c) the results, and (d) conclusions. Results are the most important. It is also highly desirable to state the number and kind of Ss, the kind of research design, and the significance levels of results" (from the *American Psychologist*, 1961, 16, 833, which may be consulted for several details).

Once you have typed up your report, put the pages together in the following order:

1. Abstract (including title, author, and affiliation)
2. Pages of text (first page includes title, author, and affiliation)
3. References (start on a new page)
4. Footnotes (start on a new page)
5. Tables (each on a separate page)
6. Figure captions (start on a new page)
7. Figures (each separately)

SOME "DO'S" AND "DON'TS"

Finally, here are a few suggestions and bits of information that did not fit conveniently into the previous sections. These matters are directed at students in the hope of improving their reports. We do not pretend to be very inclusive here, but shall simply point out items that continue to appear every year with new students. Some of them are minor, but if you learn these

arbitrary conventions now your efficiency in writing up articles later will be increased. Furthermore, the use of many of these conventions helps to facilitate communication in scientific writing.

The first thing to do before writing up a report, and this is of *great importance*, is to consult several psychological journals. Since we are concerned mainly with experimentation, you might start with the *Journal of Experimental Psychology*, although experiments are certainly reported in a large number of journals (these journals may be found in college libraries). Select several articles in these journals and study them rather thoroughly, particularly as to format, e.g., study the figures, tables, sections and subsections, and their labels; note, for example, just what words and symbols are underlined. Try to note examples of the suggestions that we have offered and particularly try to get the over-all idea of the continuity of the articles. But be prepared for the fact that you will probably not be able to understand every point in each article. Even if there are large sections that you do not understand, do not worry too much about it for this understanding will come with further learning. By the time you finish this book you will be able to understand most professional articles. Of course, some articles are extremely difficult, or in fact impossible, to understand even by specialists in the field. This is usually due to the poor quality of the write-ups.

In their own write-ups, students frequently make assertions such as: "*It is a proven fact* that left-handed people are steadier with their right hands than right-handed people are with their left hands," or "*Everyone knows* that sex education for children is good." Perhaps the main benefit to be derived by students in general from a course in experimental psychology is a healthy hostility for such statements. Before making such a statement you should have the data to back it up — you should cite a relevant reference. It is not wise to use such trite phrases as "It is a proven fact that . . ." or "Everyone knows that. . ." The use of such phrases usually indicates a lack of data on the part of the user. If you want to express one of these ideas, but lack data, the ideas still can have a place in the introductory section. They simply should be stated more tentatively, e.g., "It is possible that left-handed people are steadier with their right hands than right-handed people are with their left hands," or "An interesting question to ask is whether left-handed people . . ." Our main point, then, is that if you want to assert that something is true, make sure that you have the data (a reference) to back it up; mere opinions asserted in a positive fashion are insufficient.

Another point about writing up reports is that personal references should be kept to a minimum.¹³ For instance, students frequently say such things as

¹³Although the standard convention in scientific writing is to avoid personal references, there are those who hold that such a practice should not be sustained. The following quotation from an article entitled "Why are medical journals so dull?" states the case for this view: "... avoiding 'I' by impersonality and circumlocution leads to dullness and I would rather be thought conceited than dull. Articles are written to interest the

'I believe that the results would have turned out differently, if I had . . . ' or "It is my opinion that" Strictly speaking your audience doesn't care too much about your emotional experiences, what you believe, feel, think; they are much more interested in what data you obtained and what conclusions you can draw from those data. Rather than stating what you believe, then, you should say something like "the data indicate that. . . ." The report of an experiment falls within the context of justification rather than in the context of discovery.

Harsh or emotionally loaded phrases should also be avoided. The report of scientific work should be as divorced from emotional stimuli as possible. An example of bad writing would be: "Some psychologists believe that a few people have extrasensory perception, *while others claim it to be nonsense.*"

Misspelling occurs all too frequently in student reports. You should take the trouble to read the report over after it is written, and rewrite it if necessary. If you are not sure how to spell a word, look it up in the dictionary. Unfortunately, far too few students have acquired "the dictionary habit," a habit which is the mark of an "educated person."

When studying the format for writing up articles by referring to psychological journals, please note what sections are literally labeled. For instance, the introduction is not usually labeled as such, but the method section is always indicated, as are its subsections such as *Subjects*, *Procedure*, etc.

A few final matters are:

1. Don't list minor pieces of "apparatus," like a pencil (unless it is particularly important, as in a stylus maze).
2. Standard abbreviations that are used in the journals should be adhered to — *S* stands for "subject" and *E* stands for "experimenter" (note that they are capital, italicized letters).
3. The word "data" is plural. "Datum" is singular. Thus it is incorrect to say "that data" or "this data." Rather one should say, "those data" or "these data." Similarly "criterion" is singular and "criteria" is plural. Thus say, "This criterion may be substituted for *those* criteria." One would not be proud of himself should he say "This apples" or "Those apple."
4. There is a difference between a probability value (e.g., $P = .05$) and a percentage value (e.g., 5 per cent). Although a percentage can be changed into a probability and vice versa, one would not say that "The probability was 5 per cent," or "The per cent was .05" if he really meant 5 per cent.
5. When reporting the results of a statistical analysis, never say that "The data are (or are not) significant." Data are not significant in the technical sense. Rather, the results of your statistical analysis may indicate that there is a significant difference between your means.

reader, not to make him admire the author. Overconscientious anonymity can be overcome, as in the article by two authors which had a footnote, "Since this article was written, unfortunately one of us has died" (Asher, 1958, p. 502).

6. When you quote from an article, put quotation marks around the quote and cite a page reference.
7. Students may systematically collaborate during the data collection phase, but they should independently write up their articles and conduct their statistical analyses by themselves.
8. Make a distinction between "negative results" and "no results." Students sometimes say that they "didn't get any results," which implies (to be generous), that although they ran subjects, for some strange reason no data were collected.
9. If it is at all possible, the report should be typed, not written in long-hand. Studies have shown that students who type, get higher grades. We shall not consider why this is so, except to point out that one possible reason is that instructors have a "better unconscious mental set" in reading typed papers. In typing, set your typewriter on *double space* and leave it there for the *entire* report.

In this chapter we have examined several methods for obtaining an evidence report, the most powerful being, in turn, the experimental method, the method of systematic observation, and the clinical method. The primary difference between experimental and nonexperimental methods is that in the former the researcher produces the phenomenon of interest whereas in the latter the event is studied as it naturally occurs. We noted several steps that, given close attention, can facilitate the planning and execution of the experiment, and we discussed the write-up in some detail. To emphasize and summarize points made in writing up an experiment, we shall quote from a personal communication from Professor David A. Grant who, as we noted, has had considerable experience in reviewing manuscripts for publication:

... a paper should tell first of all what the problem is in one sentence, then a bit on why the research was done, i.e., where one proposes to contribute to our knowledge and where the findings will be relevant, etc. Then one should clearly state what was done, what was found out, and this should be complete enough so that the reader can ascertain the bases of thinking that one has found out something. Finally, in the discussion section, it should be made clear how what one has found ties in to the current knowledge and how it advances current knowledge; in this section he can also introduce qualifying clauses, and so forth.

The write-up is important — without it, the experiment might as well never have been conducted. Hence it should be written well. We cannot emphasize too strongly that you should study, in detail, articles as they appear in journals.

We now have a good overview of how to conduct an experiment. What remains is to elaborate on the many more specific topics of experimentation. The most immediate task will be to study the design and analysis procedures for an experiment in which subjects are randomly assigned to two groups.

EXPERIMENTAL DESIGN

The Case of Two Randomized Groups

You should have acquired by now a general understanding of how to conduct experiments. In Chapters One and Four we attempted to cover all the major phases of experimentation. In presenting an over-all picture of experimentation, however, it has been necessary to cover a number of steps hastily. The remaining chapters of the book will consist of attempts to fill in these relatively neglected areas. But we should try never to lose sight of how the steps on which we momentarily concentrate fit into the general picture of designing and conducting an experiment.

We shall now focus on the phase of experimentation that concerns the selection of a design. Although there are a number of designs available to the experimenter, we have thus far limited our consideration to one that involves only two groups. In this chapter we shall discuss this type of design more thoroughly. Since the “two-groups” design is basic in psychology, an understanding of it will form a sound foundation from which we can move to more complex (though not necessarily more difficult to comprehend) designs.

A GENERAL ORIENTATION

The "two-groups" design may more completely be referred to as the "two-randomized-groups design." To summarize briefly what has been said about this design, let us recall that the experimenter defines an independent variable that he seeks to vary in (at least) two ways. The two values that he assigns to the independent variable may be referred to as two "conditions," "treatments," or "methods." He then seeks to determine whether these two conditions differentially affect his dependent variable. The general procedure that he follows to answer this question may be summarized as follows. First, he defines a certain population about which he wishes to make a statement. Then he randomly selects a sample of subjects to study. Since that sample has been drawn randomly from the population it may be assumed to be representative of the population. Thus what is observed in the sample is used to make inferences that the same holds for the population. Let us assume that the population is defined as all students in a certain university. They may number 6,000. We decide that our sample shall be 60. One reasonable method for selecting this sample would be to obtain an alphabetical list of the 6,000 students. Then randomly select one name from the first one hundred and take every 100th student on that list after that. On the assumption that all 60 students will cooperate, we now have our sample. It has been specified that we will study two conditions in our experiment. To assign a separate group of subjects to each condition, we must divide the 60 subjects into two groups. Again, any method that would assure that the subjects are randomly assigned to the two groups would suffice. Let us say that we write the name of each subject on a separate slip of paper and place all 60 pieces of paper in a hat. We may then decide that the first name drawn would be assigned to the first group, the second to the second group, the third to the first group, etc. In this manner we would end up with two groups, each with thirty subjects. A simple flip of a coin would then tell us which is to be the experimental group, and which the control group. The reason that this is called the "two-randomized-groups design" is now quite apparent: Subjects are *randomly* assigned to *two* groups.

A basic and important presupposition made in any type of design is that the means (averages) of the groups do not differ significantly at the start of the experiment. In a two-groups design the two values of the independent variable are then respectively administered to the two groups. For example, some positive amount of the independent variable might be administered to the experimental group, while a zero amount is administered to the control group. Scores of all subjects on the dependent variable are then recorded and subjected to statistical analysis. If the appropriate statistical test indicates that the two groups are significantly different (on the dependent variable

scores), it may be concluded that this difference is due to the variation of the independent variable — assuming that the proper experimental controls have been in effect, it may be concluded that the two values of the independent variable are effective in producing the differences in the dependent variable.

“EQUALITY” OF GROUPS THROUGH RANDOMIZATION

Now, by randomly assigning the 60 subjects to two groups, we said, it is highly reasonable to assume that the two groups are essentially equal; but approximately equivalent with respect to what? The answer might be that the groups as wholes are equivalent in all respects. And such an answer is easy to defend, assuming that the randomization has been properly carried out. In any given experiment, however, we are not interested in comparing the two groups in all respects. Rather, we want them to be equal on those factors that might affect our dependent variable. Suppose the dependent variable concerns the rate at which a person learns a task that involves visual abilities. In this case we would want the two groups to be equivalent at least with respect to intelligence and visual acuity. More particularly, we would want the means of intelligence and visual acuity scores to be essentially the same. For both of these factors are likely to influence scores on our dependent variable.

Students frequently criticize the randomized-groups design by pointing out that “by chance” (i.e. due to random fluctuations) we could end up with two very unequal groups. It is possible, they say, that one group would be considerably more intelligent, on the average, than the other group, that is, that one group would have a higher mean intelligence score. Even though such an event is indeed possible, it is unlikely, particularly if a large number of subjects is used in both groups. For it can be demonstrated that the larger the number of subjects randomly assigned to the two groups, the closer their means come to each other. Hence, although with a small number of subjects it is unlikely that the means of the two groups will differ to any great extent, it is more likely than if the number of subjects is large. The lesson should be clear: If you wish to reduce the difference in the means of the two groups, use a large number of subjects.¹

Even with a comparatively large number of subjects it is still possible, though unlikely, that the means of the groups will differ considerably due to random fluctuations. Suppose, for example, that we have drawn a sample of sixteen subjects and assigned them to two groups. Now, if we measured their intelligence, it is possible that we would obtain a mean intelligence quotient

¹In making this point we are ignoring the distributions of the scores. Hence, the matter is not quite as simple as we have made it, but the main point is sound.

of 100 for one group and a mean of 116 for the second group. However, by using appropriate statistical techniques we can determine that such an event should occur by chance less than about five times out of 100. If we ran the experiment 100 times, and assigned subjects to two groups at random in each experiment, a difference between the groups of 16 IQ points (e.g., 116-100) or more should occur by chance in only about five of the experiments. Differences between the two groups of less than sixteen IQ points should occur more frequently. And differences between the two groups of 24 points or more should occur less than one time in 100 experiments, on the average. Most frequently, then, there should be only a small difference between the two groups.

"But," the skeptical student continues, "suppose that in the particular experiment that I am conducting (I don't care about the other 95 or 99 experiments) I *do* by chance divide my subjects into two groups of widely differing ability. I would think that the group with the mean IQ of 116 would have a higher mean score on the dependent variable than does the other group, *regardless of the effect of the independent variable*. I (the experimenter) would then conclude that the independent variable is effective, when, in fact, it isn't."

One cannot help but be impressed by such a convincing attack, but retreat at this point would be premature, for there are still several weapons that can be brought into the battle. First, if one has doubts as to the equivalence of his two groups, he can compute their scores on certain variables to see how their means actually compare. Thus, in the above example, we could measure the subjects' IQ's and visual acuity, compute the means for both groups, and compare the scores to see if there is much difference. If there is little difference, we know that our random assignment has been at least fairly successful. This laborious and generally unnecessary precaution actually has been taken in a number of experiments.²

"But," the student continues tenaciously, "suppose I find that there is a sizeable difference, and furthermore, suppose that I determine this only after all data have been collected. My experiment would be invalidated." Yet, there is hope. In this case we could use a statistical technique that allows us to equate the two groups with respect to intelligence. That is, we could "correct" for the difference between the two groups and determine whether they differ on the dependent variable for a reason that cannot be attributed to the difference of intelligence. Put another way, we could statistically equate the two groups on intelligence so that differences on this extraneous variable would not differentially affect the dependent variable scores. Fur-

²In an experiment on rifle marksmanship, for instance, it was determined that four groups did not differ significantly on the following extraneous variables: previous firing experience, left or right handedness, visual acuity, intelligence, or educational level (McGuigan and MacCaslin, 1955, a).

ther consideration of this statistical technique (known as the analysis of covariance) would be premature here (see p. 360).

"Excellent," the student persists, "but suppose the two groups differ in some respect for which we have no measure, and that this difference will sizeably influence scores on the dependent variable. I now understand that we can probably 'correct' for the difference between the two groups on factors such as intelligence and visual acuity, because these are easily measurable variables. But what if the groups differ on some factor that we cannot measure or do not think to measure? In this case we would be totally unaware of the difference and draw illegitimate conclusions from our data."

"You," we say to the student, secretly admiring his demanding perseverance, "have now put us in such an unlikely position that we need not worry about its occurrence. Nevertheless, it is possible, just as it is possible that you will be hit by a car today while crossing the street. And, if there is some factor for which we cannot make a 'correction,' the experiment might well result in erroneous conclusions." The only point we can refer to here is one of the general features of the scientific enterprise: Science is self-correcting. Thus, if any given experiment leads to a false conclusion, and if the conclusion has any importance at all for psychology, an inconsistency between the results of the invalid experiment and additional data from a later experiment will become apparent. The existence of this problem will then lead to a solution, which, in this case, will be a matter of discarding the incorrect conclusion.

STATISTICAL ANALYSIS OF THE TWO-RANDOMIZED-GROUPS DESIGN³

Now, the matter for us to discuss here concerns a return to a question that was briefly considered in Chapter One, where we posed the following problem: After the experimenter has collected his data on the dependent variable, he wishes to determine whether one group is superior to the other. His hypothesis may predict that the experimental group will have a higher mean than the control group. The first step in testing the hypothesis would be to compute the mean scores on the dependent variable for the two groups. It might be found that the experimental group has a higher mean score than the control group — say that the experimental group has a mean score of 40, while the control group has one of 35. Assuming that the higher the score the better the performance, can we conclude that this five-point difference is significant? Or is it merely the result of random fluctuations, of experimental error? To answer this question, we said, we must apply a

³Before beginning this section the conscientious student might want to read the first section of Chapter Fifteen, "Concerning the Accuracy of the Data Analysis."

statistical test. Let us now consider one statistical test that is frequently used to answer this question.

The statistical test to which we refer is known as the "*t* test" (note that a lower-case "*t*" is used to denote this test, not a capital "*T*," which has another denotation in statistics).⁴ The first step in computing a *t*-test value is the computation of the means of the dependent variable scores of the two groups concerned. The equation for computing a mean (symbolized \bar{X}) is

$$(5.1) \quad \bar{X} = \frac{\Sigma X}{n}$$

The only unusual symbol in Equation (5.1) is Σ , the capital Greek letter sigma. Σ may be interpreted as "sum of." It is a summation sign and simply instructs you to add whatever is to the right of it.⁵ In this case the letter X is to the right of sigma so we must now find out what values X stands for and add them. Here, X merely indicates the score that we obtained for each subject. Suppose, for instance, that we give a test to a class of five students, with these resulting scores:

	X
Joan	100
Constance	100
Richard	80
Lillian	70
Joe	60

To compute ΣX we merely need to add the X scores. In this way we find that $\Sigma X = 100 + 100 + 80 + 70 + 60 = 410$. The n in Equation (5.1) stands for the number of subjects in the group. In this example, then, $n = 5$. Thus, to compute \bar{X} we simply substitute 410 for ΣX , 5 for n in Equation (5.1), and then divide n into ΣX :

$$\bar{X} = \frac{410}{5} = 82.00$$

Thus the mean score of the group of 5 students who took the particular test is 82.00.

⁴Before our first discussion of the statistical analysis of an experimental design, it is well to point out that the statistical tests (such as the *t* test) are conducted on the supposition that certain statistical assumptions are satisfied. Since the assumptions for all the statistical tests discussed in this book are similar, it is more economical to discuss them together after all of our designs have been considered (p. 354). The instructor or student who so wishes, of course, may immediately integrate this topic with the discussion of the statistical tests.

⁵More precisely, Σ instructs you to add all the values of the symbols that are to its right, values that were obtained from your sample.

Let us now turn to an equation for computing t :

$$(5.2) \quad t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{SS_1 + SS_2}{(n_1 - 1) + (n_2 - 1)}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

While this equation may look forbidding to the statistically naive, such an impression should be short-lived for t is actually rather simple to compute. To illustrate, consider the fascinating (and somewhat controversial) results of one of a series of experiments on RNA in the brain. RNA (Ribonucleic acid) has been implicated in the process of memory storage, though its precise role has yet to be established. More or less grabbing the bull by the horns, Babich, Jacobson, Bubash and Jacobson (1965) trained a group of eight rats to approach the food cup in a Skinner Box every time a click was sounded. The animals rarely or never approached the food cup when the click was absent. On the day after this training was completed, the animals were sacrificed, their brains were removed, and RNA was extracted from a selected portion. RNA was also extracted from the brains of nine untrained rats. Approximately eight hours after extraction, the RNA from each of the rats, trained and untrained, was injected into live, untrained rats. Hence, seventeen live rats were injected with RNA: eight of them (experimental group) received RNA from trained rats and nine (control group) received RNA from untrained rats. Both groups were then tested in a Skinner Box by presenting a click for 25 times, and the number of times that they approached the food cup was counted (various controls were used, but they need not concern us here). The hypothesis, amazing as it might sound, was to the effect that memory storage could be passed on by means of injections of RNA or associated substances. It was therefore predicted that the experimental group would approach the food cup more often during the test trials than the control group. The number of times that each rat approached the food cup during the 25 test trials is presented in Table 5.1 (one rat from each group was discarded because it "froze" in the test situation).

Table 5.1. *Number of Food Cup Approaches per Animal During 25 Test Trials.*

Subject Number	GROUP 1	Subject Number	GROUP 2
	Experimental Rats \bar{X}_1		Control Rats \bar{X}_2
1	1	8	0
2	3	9	0
3	7	10	0
4	8	11	1
5	9	12	1
6	10	13	1
7	10	14	2
	$\Sigma X_1 = 48$	15	3
			$\Sigma X_2 = 8$

We now seek to obtain an evidence report, i.e., a summary statement of the findings of the study. This evidence report, then, will tell us whether the hypothesis is probably true or false. The first step is to compute the means of the two groups. Note that subscripts have been used in Equation (5.2) to indicate which group the various values are for. In this case \bar{X}_1 stands for the mean of Group 1 (the experimental group), and \bar{X}_2 for the mean of Group 2 (the control group). In like manner SS_1 and SS_2 stand for what is called the *sum of squares* for Groups 1 and 2 respectively; and n_1 and n_2 are the respective number of subjects in the two groups. We can now determine that $\Sigma X_1 = 48$, while $\Sigma X_2 = 8$. Since the number of subjects in Group 1 is seven, we note that $n_1 = 7$. The mean for Group 1 (i.e., \bar{X}_1) may now be determined by substitution in Equation (5.1):⁶

$$\bar{X}_1 = \frac{48}{7} = 6.86$$

And similarly for Group 2 (n_2 is 8):

$$\bar{X}_2 = \frac{8}{8} = 1.00$$

We now need to compute the sum of squares (a term that will be extensively used in later chapters) for each group. The equation for the sum of squares is:

$$(5.3) \quad SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

Equation (5.3) contains two terms with which we are already familiar, viz., n and ΣX . The other term is ΣX^2 and it instructs us to add the squares of all the values for a given group. Thus, to compute ΣX^2 for Group 1 we should square the value for the first subject, add it to the square of the score for the second subject, add both of these values to the square of the score for the third subject, etc. Squaring the scores for the subjects in both groups and summing them we obtain:

⁶In your computations you would be wise to pay attention to the significant figures, an indication of the accuracy of your measurements and computations. To determine the accuracy of a measurement you count the number of digits, e.g., 21 is correct to two significant figures, 1.2 to two significant figures, .012 to two significant figures, and 1.456 to four significant figures. The final value of statistics, like a mean or standard deviation, should be rounded off to one more significant figure than for the raw data. Intermediate calculations for the t test can be safely performed by carrying three more digits than the data.

Subject Number	GROUP 1		Subject Number	GROUP 2	
	Experimental	Rat.		Control	Rats
	X_1	X_1^2		X_1	X_2^2
1	1	1	8	0	0
2	3	9	9	0	0
3	7	49	10	0	0
4	8	64	11	1	1
5	9	81	12	1	1
6	10	100	13	1	1
7	10	100	14	2	4
		$\Sigma X_1^2 = 404$	15	3	9
				$\Sigma X_2^2 = 16$	

One frequent error by students should be pointed out as a precaution. That is that ΣX^2 is not the square of ΣX . That is $(\Sigma X)^2$ is not equal to ΣX^2 . For instance, the $\Sigma X_1 = 48$. The square of this value is $(\Sigma X_1)^2 = 2304$, whereas $\Sigma X_1^2 = 404$.

We are now in a position to substitute the appropriate values into Equation (5.3) and compute the sum of squares for each group. We know that, for Group 1, $\Sigma X_1 = 48$, that $\Sigma X_1^2 = 404$, and that $n_1 = 7$. Hence:

$$\begin{aligned} SS_1 &= 404 - \frac{(48)^2}{7} = 404 - \frac{(48 \cdot 48)}{7} \\ &= 404 - \frac{(2304)}{7} = 404.000 - 329.143 = 74.857 \end{aligned}$$

And similarly, the values to compute the sum of squares for Group 2 are:

$$\Sigma X_2 = 8$$

$$\Sigma X_2^2 = 16$$

$$n_2 = 8$$

Therefore:

$$SS_2 = 16 - \frac{(8)^2}{8} = 16 - \frac{64}{8} = 8.000$$

We now have all of the values required by Equation (5.2), and therefore can immediately compute the value of t for this experiment.

To summarize them:

$$\bar{X}_1 = 6.86$$

$$\bar{X}_2 = 1.00$$

$$n_1 = 7$$

$$n_2 = 8$$

$$SS_1 = 74.857$$

$$SS_2 = 8.000$$

And substituting these values in Equation (5.2) we obtain:

$$t = \frac{6.86 - 1.00}{\sqrt{\left(\frac{74.857 + 8.000}{(7-1) + (8-1)}\right)\left(\frac{1}{7} + \frac{1}{8}\right)}}$$

We now need to go through the following steps in computing t :

1. Obtain the difference between the means: $6.86 - 1.00 = 5.86$
2. Add $SS_1 + SS_2$: $74.857 + 8.000 = 82.857$
3. Compute $n_1 - 1$: $7 - 1 = 6$
4. Compute $n_2 - 1$: $8 - 1 = 7$
5. Add $\frac{1}{n_1} + \frac{1}{n_2} = \frac{1}{7} + \frac{1}{8} = \frac{8}{56} + \frac{7}{56} = \frac{15}{56}$

Substituting these values:

$$t = \frac{5.86}{\sqrt{\left(\frac{82.857}{6+7}\right)\left(\frac{15}{56}\right)}}$$

In the next stage divide the two denominators (13 and 56) into their respective numerators (82.857 and 15):

$$t = \frac{5.86}{\sqrt{(6.374)(.2679)}}$$

then multiply the values in the denominator:

$$t = \frac{5.86}{\sqrt{1.708}}$$

The next step is to find the square root of 1.708. This may be obtained from page 366 in the Appendix, and is found to be 1.307. Dividing as indicated we find t to be:

$$t = \frac{5.86}{1.307} = 4.48$$

Although the computation of t is straightforward, the beginning student is likely to make an error in its computation. The error is generally not one of failing to follow the procedure, but one of a computational nature (dividing incorrectly, failing to square terms properly, mistakes in addition). A great deal of care must be taken in statistical work; each step of the computation should be checked in an effort to eliminate errors. As an aid to the student in learning to compute t , a number of exercises are provided at the end of this chapter. Work all of these exercises and make sure that your answers are correct.

One point in the computation of t needs to be clarified. In the numerator we have indicated that \bar{X}_2 should be subtracted from \bar{X}_1 . Actually we are

conducting what is known as a "two-tailed test." You need not be concerned about this term here, but the important point for you to observe is that we are interested in the absolute difference between the means. Hence $\bar{X}_1 - \bar{X}_2$ is appropriate if \bar{X}_1 is greater than \bar{X}_2 (i.e., if $\bar{X}_1 > \bar{X}_2$). But if in your experiment you find that \bar{X}_2 is greater than \bar{X}_1 ($\bar{X}_2 > \bar{X}_1$) then you merely subtract \bar{X}_1 from \bar{X}_2 , i.e., Equation (5.2) would have as its numerator $\bar{X}_2 - \bar{X}_1$.

We might also note that the value under the square root sign is always positive. If it is negative in your computation, go through your work to find the error.

The reason we want to obtain a value of t , we said, is to decide whether the difference between the means of two groups is the result of random fluctuations or whether it is a significant difference. But several additional matters must also be discussed in relation to this difference. The first is a consideration of what is known as the "null hypothesis," a concept that it is vital to understand. The null hypothesis that is generally used in psychological experimentation, roughly, states that there is no difference between two groups. Since we wish to contrast the two groups by comparing their means on the dependent variable, we may more precisely state that there is no difference between the population means on the dependent variable of the two groups.

The null hypothesis, let it be emphasized, states that there is no difference between the *population* means.⁷ The reason we conduct an experiment is to make statements about populations — to determine whether the population means of our two groups differ. In a sense this may be stated otherwise as follows: we want to know whether the *true* means of our groups differ (where the true mean is taken as the population mean). Now, of course, we cannot study the population in its entirety. Rather, the way to determine whether or not the true (population) means differ is by comparing the means obtained for our two sample groups. We do this by subtracting one sample mean from the other, as specified in the numerator of Equation (5.2). This difference will almost certainly not be zero; but it will be some positive amount, the value of which may be quite small or quite large. If the difference between our sample means is quite small, we would be inclined to conclude that the difference is due to chance.

On the other hand, if the difference is large, then we might say that the difference is too large to be due to random fluctuations alone. The null hypothesis asserts that the difference between the population means is zero.

⁷A symbolic statement of the null hypothesis would be $\mu_1 - \mu_2 = 0$ (μ is the Greek letter mu). Here μ_1 is the population mean for Group 1 and μ_2 is the population mean for Group 2. If the difference between the sample means ($\bar{X}_1 - \bar{X}_2$) is small, then we are likely to infer that there is no difference between the population means; thus, that $\mu_1 - \mu_2 = 0$. On the other hand if $\bar{X}_1 - \bar{X}_2$ is large, then the null hypothesis that $\mu_1 - \mu_2 = 0$ is probably not true.

In effect it says that any difference between the means of the groups in your sample is merely due to random fluctuations and thus can be accounted for in terms of experimental error. If we find that the difference between the means of our groups is small, then it is likely that the difference is the result of random fluctuations and that the null hypothesis is reasonable. But if our groups differ considerably, then the difference is probably too large to be due to random fluctuations alone, and the null hypothesis is not tenable in that particular case.

The question now is how small the difference must be between \bar{X}_1 and \bar{X}_2 before we can say that it is due to random fluctuations of the means. Then, too, how large must the difference be before we can say that it is *not* due to random fluctuations alone? The latter question can be answered by the value of t ; if t is sufficiently large, we can say that the difference between the two groups is too large to be attributed solely to random fluctuations — too large to be accounted for by experimental error. And to determine how large “sufficiently large” is we may consult the table of t . But before doing this, there is one additional value that we must compute — the degrees of freedom (df). The degrees of freedom available for the t test are a function of the number of subjects in the experiment. More specifically, $df = N - 2$.⁸ And N is the number of subjects in one group (n_1) plus the number of subjects in the other group (n_2). Hence, in our example we have:

$$N = n_1 + n_2 \quad \text{i.e.,} \quad N = 7 + 8 = 15$$

therefore:
$$df = 15 - 2 = 13$$

To determine whether our t is significant, let us now turn to a table of t (Table 5.2 on p. 108) armed with two values: $t = 4.48$ and $df = 13$. The table of t is organized around two values: a column labeled “ df ” and a row labeled “ P ” (for probability). The df column is on the extreme left, and the P row runs across the top of the table. Values of t are the numbers that complete the table. Our general purpose here is to find out what P value is associated with a specific value of t and df . To do this we must first run down the df column until we arrive at our specific value of df ; in this case, 13 df . We then read across the row that is marked 13 df . This row contains a number of possible values of t : 0.128, 0.259, 0.394, etc. We must read across this row until we come to a value of t that is close to our particular value — in this case, 4.48. But the largest value of t in this row is 3.012; this value, then, is the closest match we can make to 4.48. So, we read up the column that contains 3.012 to determine what value of P is associated with it — in this case, 0.01.

⁸This equation for computing df is only for the application of the t -test to two randomized groups. We shall use other equations for df when considering additional statistical tests.

Let us make a general observation; the larger the t , the smaller the P . For example, with 13 df a t of 0.128 has a P of 0.9 associated with it, while with the same df a t of 1.771 has a P of 0.1. From this observation and our study of the tabled values of t and P we can conclude that if a t of 3.012 has a P of 0.01, any t larger than 3.012 must have a smaller P than .01. It is sufficient for our purposes simply to note this fact without attempting to make it any more precise.

The next step is to interpret the fact that a t of 4.48 has a P of less than 0.01 ($P < 0.01$) associated with it. This finding indicates that a difference between means of the two groups of the size that were obtained has a probability of less than 0.01; i.e., that a difference between the means of this size may be expected less than one time in a hundred by chance ($.01 = 1/100$). Put another way, if the experiment had been conducted one hundred times, by chance we would expect a difference of this size to occur once, provided the null hypothesis is true. This, we must all agree, is a most unlikely occurrence. It is so unreasonable, in fact, to think that such a large difference could have occurred by chance on the very first of the hypothetical one hundred experiments that we prefer to reject "chance" as the only explanation. We therefore choose to reject our null hypothesis. That is, we refuse to regard it as reasonable that the real difference between the means of the two groups is zero when we have obtained such a large difference in sample means, as indicated by the respective values, in this case, of 6.86 and 1.00. But if a difference of this size cannot be attributed to chance, alone, what reason can we give for it? We assume that all the proper safeguards of experimentation have been observed in obtaining these results, and that the groups therefore differed only in the respect that each was administered a different experimental treatment. It seems reasonable to assert, then, that the reason the two groups differed is that they received different values of the independent variable. This leads to the further conclusion that the independent variable is effective in influencing scores on the dependent variable; and this is precisely the purpose of the experiment.

There are still a number of questions about this procedure that need to be answered. For instance, we said that before we conduct an experiment it is unlikely that we would (by chance) obtain a P of .01 for our t . How small may P be and still be considered sufficiently likely to occur by chance alone? That is, how small must P be before we can reject the null hypothesis? For example, with 13 df , if we had obtained a value of 1.80 for t , we find that the corresponding value of P is less than .10. This would imply that such a difference between the two group means could be expected by chance less than ten times out of 100. Now, is this sufficiently unlikely that we can reject the null hypothesis? Again, consider a t of 2.20. The corresponding P is less than 0.05. Can we reject the null hypothesis on the basis of this size of P ? What if we had obtained a t of 0.90, with a corresponding P of less

Table 5.2.* Table of t .

df	\mathcal{P}	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01
1		0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2		0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3		0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4		0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5		0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6		0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7		0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8		0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9		0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10		0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11		0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12		0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13		0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14		0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15		0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16		0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17		0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18		0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19		0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20		0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21		0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22		0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23		0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24		0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25		0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26		0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27		0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28		0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29		0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30		0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
∞		0.12566	0.25335	0.38532	0.52440	0.67449	0.84162	1.03643	1.28155	1.64485	1.95996	2.32634	2.57582

*Table 5.2 is reprinted from Table IV of Fisher: *Statistical Methods for Research Workers*, published by Oliver and Boyd Ltd., Edinburgh, by permission of the author and publishers.

than 0.40; a difference of this size may be expected 40 times out of 100 by chance. Is this P sufficiently small to allow us to reject the null hypothesis? In short, the question is this: what value of P is small enough to allow us to reject the null hypothesis? Unfortunately, there is no simple answer to this question, for it depends on a number of things. The best we can say here is that the experimenter may set any value of P as the cut-off point. Thus he may say that: "If the value of t that I obtain has a P of less than 0.50, I will reject my null hypothesis." Similarly, he may set P at 0.01, 0.05, 0.30, or even 0.90 if he wishes. There is only one requirement that he must satisfy in setting the value of P : he must set it *before* he conducts his experiment. The reason for this is that, for a proper test of the null hypothesis, the experimenter should not be influenced by the particular nature of his data. For example, it would be inappropriate to run a t test and determine P to be 0.06, and then decide that if P is 0.06 to reject the null hypothesis. Such an experimenter would be inclined never to fail to reject the null hypothesis, for the criterion (the value of P) for rejecting it would be determined by the value of P actually obtained. An extreme case of this would be a person who obtains a P of 0.90, and then sets 0.90 as his criterion. The sterility of such a decision is apparent, for a difference between his groups of the size that he obtained would be expected by chance 90 times out of 100. Obviously, it would be unreasonable to reject the null hypothesis with such a large P , and such an experimenter would almost surely be committing an error, i.e., he would be rejecting the null hypothesis when in fact it should not be rejected.

The P that an experimenter sets, then is totally arbitrary. He can vary it with the particular experiment that he is conducting. For some problems it is important to have an extremely small P , for others a larger one is appropriate. Although the actual decision is arbitrary, there are a number of important considerations that will help the experimenter in arriving at his decision. The interested student will find such matters discussed in elementary statistics courses. Suffice it to say here that for general psychological experimentation a standard value of P is accepted: 0.05. Thus, unless the experimenter specifies otherwise, it is generally understood that he has set a P of .05 before conducting his experiment.⁹

Let us now apply the above considerations to our example. The hypothesis was to the effect that the sample of experimental animals should approach the food cup significantly more frequently during the test than the controls. It was found that the mean scores for the two groups were 6.86 and 1.00, respectively. Furthermore, we found that the t -test yielded a value of 4.48, which, with 13 df , had a P of less than 0.01. Assuming the conventional value of 0.05 for P as the criterion of whether or not to reject the null hypothesis, the

⁹The value of P set as the criterion of rejecting or failing to reject the null hypothesis is known as the *level of significance*. Sometimes people erroneously call the level of significance "the confidence level" or "the level of confidence."

value of less than 0.01 causes us to reject the null hypothesis. That is, we assert that there is a true difference between our two groups. Furthermore, we observe that the direction of the difference is that specified by the (empirical) hypothesis, i.e., the hypothesis predicted that the scores for the experimental rats would be higher than for the control animals. Since the scores are of the nature predicted by the hypothesis (and significantly so), we may conclude that the hypothesis is confirmed.^{10,11}

The following general rules may now be stated: *If the empirical hypothesis predicts that there will be a difference between two groups, and if the null hypothesis is rejected, and if the difference between the two groups is in the direction specified by the empirical hypothesis, then it may be concluded that the empirical hypothesis is confirmed.* Thus, there are two cases in which the empirical hypothesis would not be confirmed: first, if the null hypothesis were not rejected; and second, if it were rejected, but the difference between the two groups were in the opposite direction specified by the empirical hypothesis. To illustrate these latter possibilities, let us assume that we actually obtained a t of 1.40 (which you can see has a P value greater than .05). We fail to reject the null hypothesis, and accordingly fail to confirm the empirical hypothesis. On the other hand, assume that we obtain a t of 2.40 ($P < 0.05$), but that the mean score for the controls is higher than that for the experimental rats. In this case we reject the null hypothesis, but fail to confirm the empirical hypothesis.

For further practice, and for an experiment in which the values are more typical, let us consider the following study. In this experiment the role of "contact comfort" in the development of "mother love" in dogs was studied (Igel and Calvin, 1960). The suggestion that the sense of touch is important in this regard is an old one. In 1868 Bain, for instance, said that ". . . touch is the fundamental and generic sense The soft, warm touch, if not a first-class influence, is at least an approach to that. The combined power of soft contact and warmth amounts to considerable pitch of massive pleasure In a word, our love pleasures begin and end in sensual contact It seems to me that there must be at the (parental instinct's) foundation that

¹⁰One of the assumptions of parametric statistics such as the t -test, as you will see in Chapter Fifteen, is that the variances of the groups are homogeneous ("equal"). In this experiment they are not, viz., the variance for the experimental group is 12.47 and that for the control group is 1.14. One alternative to a parametric analysis is to conduct a nonparametric test, as was done in the original report of the experiment. The point that we make later, however, is well illustrated by this example, i.e., that parametric tests are remarkably robust in that major deviations from their basic assumptions can be tolerated. Here the same conclusions follow from the t -test and from the Mann-Whitney U test, a nonparametric test.

¹¹To emphasize the controversial nature of the findings reported here (and also in another experiment in Chapter 9), we need only observe the results reported by other experimenters. Batkin, Woodward, Cole, and Hall (1966) report a similar effect when using carp, but Byrne (1966) cites 18 experiments on this problem that yielded negative results.

intense pleasure in the embrace of the young which we find characterize the parental feeling throughout" (James, 1952, p. 809).

In a series of studies Harlow and his associates have found that "... contact comfort is a variable of overwhelming importance in the development of affectional responses, whereas lactation is a variable of negligible importance" (Harlow and Zimmerman, 1958, p. 503). The purpose of the Igel and Calvin study was to determine whether Harlow's findings were species-specific, or whether they could be generalized to species other than the monkey. The hypothesis, thus, was to the effect that "puppies will spend more time with cloth than with wire mothers, even though they are fed by both types." Among the experimental conditions was one in which one group of puppies was raised and fed on a wire surrogate mother (Fig. 5.1) while the other

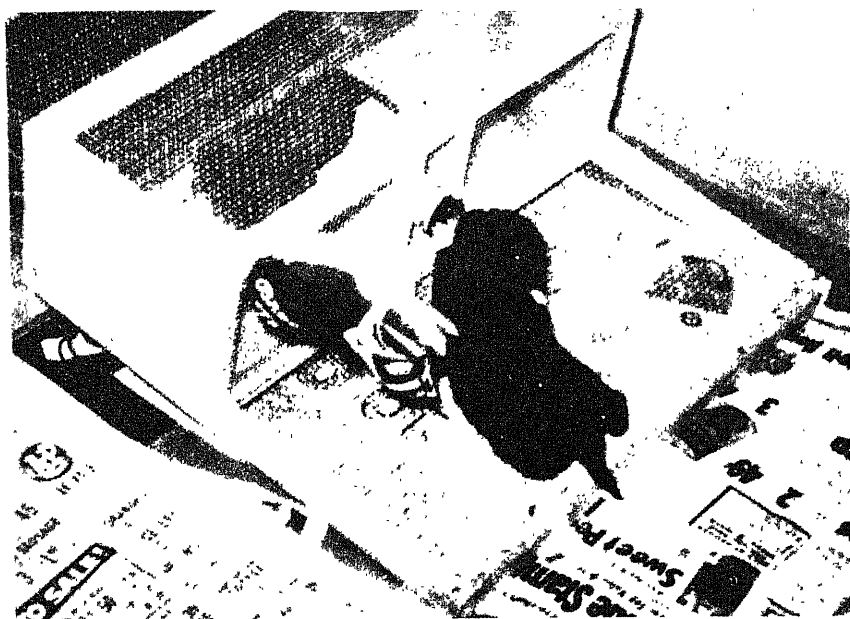


FIGURE 5.1.

S nursing from a wire surrogate mother.

group was similarly raised with a terry cloth mother (Fig. 5.2). The latter obviously provides more contact comfort; but would a puppy learn to love even a wire mother if, as others have held, she feeds him? The index of affection was the amount of time that each puppy spent with his mother surrogate. Hence, the more a puppy "loves" his mother, the longer the amount of time he would spend with it. Amount of contact time was automatically recorded for each day from the time the pups had their eyes open

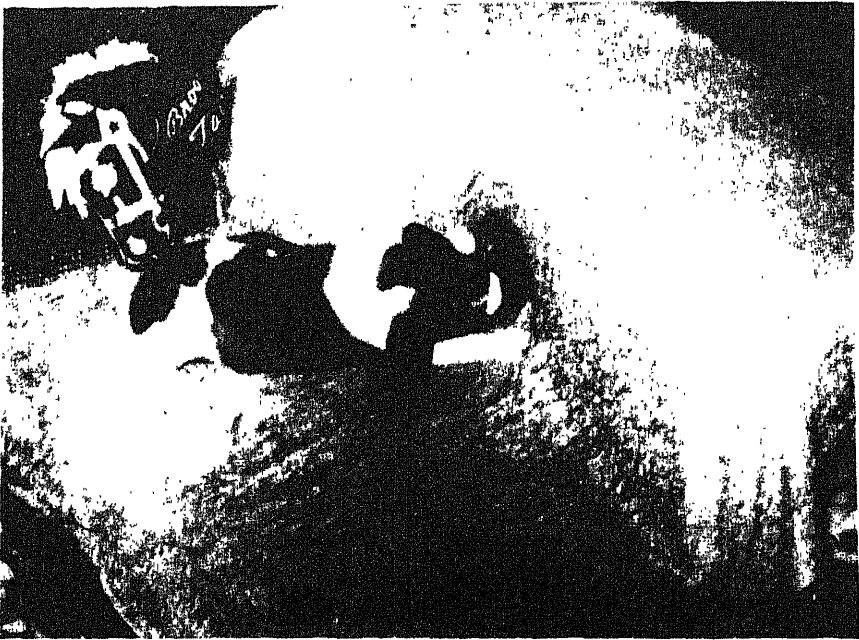


FIGURE 5.2.
S nursing from a cloth surrogate mother.

(11 days old) to 31 days old. In the condition that we are considering, two puppies were randomly assigned to a wire mother and two to a cloth mother. The dependent variable scores are presented in Table 5.3.

Table 5.3. *Mean Number of Minutes Per Day Spent with Surrogate Mothers.*

<i>Subject Number</i>	<i>GROUP 1 Wire Mother</i>	<i>Subject Number</i>	<i>GROUP 2 Cloth Mother</i>
1	61.2	3	836.5
2	82.7	4	722.3
	$\Sigma X_1 = 143.9$		$\Sigma X_2 = 1558.8$
	$\bar{X}_1 = 71.95$		$\bar{X}_2 = 779.40$

Computing the necessary values for the sums of squares Equation (5.3) we find:

<i>Group 1</i>	<i>Group 2</i>
$\Sigma X_1 = 143.9$	$\Sigma X_2 = 1558.8$
$\Sigma X_1^2 = 10,584.73$	$\Sigma X_2^2 = 1,221,449.54$
$n_1 = 2$	$n_2 = 2$

The sums of squares for Group 1 is:

$$\begin{aligned} SS_1 &= 10,584.73 - \frac{(143.9)^2}{2} = 10,584.73 - \frac{(20707.21)}{2} \\ &= 10,584.73 - 10,353.61 = 231.12 \end{aligned}$$

And for Group 2:

$$\begin{aligned} SS_2 &= 1,221,449.54 - \frac{(1558.8)^2}{2} \\ &= 1,221,449.54 - 1,214,928.72 = 6520.82 \end{aligned}$$

We now have all the necessary values for Equation (5.2) and substitute them as follows:

$$t = \frac{779.40 - 71.95}{\sqrt{\left(\frac{231.12 + 6520.82}{(2-1) + (2-1)}\right)\left(\frac{1}{2} + \frac{1}{2}\right)}}$$

Performing the indicated operations:

$$\begin{aligned} t &= \frac{707.45}{\sqrt{\left(\frac{6751.94}{2}\right)(1)}} = \frac{707.45}{\sqrt{3375.97}} \\ &= \frac{707.45}{58.103} = 12.18 \end{aligned}$$

We thus find that $t = 12.18$. With two degrees of freedom ($N - 2 = 4 - 2 = 2$) we refer to our table of t , and find that the P associated with it is less than 0.01. Since this P is less than our conventional criterion of 0.05, we reject the null hypothesis and conclude that there is a difference between the groups. On the assumption that necessary controls were satisfactorily effected, we may conclude that the independent variable was effective. That is, that the reason the puppies spent more time with the cloth mother was because of the difference in texture. These findings thus supported Harlow's contention that the strength of the affectional response is a function of the texture of the mother's "skin."

At this point it may be beneficial to discuss further the null hypothesis. This hypothesis is a statistical hypothesis and is set up for the purpose of attempting to disprove it. Our null hypothesis asserts that there is no difference between the population means of our two groups; we seek to determine that it is false, that there is a difference between the means. Hence, if it is disproven in a properly conducted experiment, we can conclude that there is a difference between our two groups and furthermore that this difference is due to variation of the independent variable. If we cannot disprove the null hypothesis, then we cannot assert that there is a difference between the two groups; variation of our independent variable is not effective.

Incidentally, two characteristics of the null hypothesis may be illustrated by reference to the word *null*. First, it may be noted that “null” derives from the Latin *nullus*, “not any.” Hence, a purpose of experimentation is to determine whether the difference between the experimental result and that specified by the null hypothesis is null. If there is a “null” discrepancy between the experimental results and the null hypothesis, then we do not reject the null hypothesis. If, however, the discrepancy is positive, is not “null,” then we reject it.

The second association is with the word “nullify.” In this case the null hypothesis is the one that we seek to nullify. If we can nullify (reject) it, then we may conclude that the independent variable is effective. But if we fail to nullify it, we cannot assert that the independent variable is effective.¹²

Let us now summarize each major step that we have gone through in testing an empirical hypothesis. For this purpose you might design a study to compare the amount of anxiety of majors in different college departments.

1. State the hypothesis, e.g., “If the anxiety scores of English and Psychology students are measured, the Psychology students will have the higher scores.”

2. The experiment is designed according to the procedures outlined in Chapter Four, e.g., “anxiety” is operationally defined [such as scores on the Manifest Anxiety Scale, Taylor, 1953], samples from each population are drawn, etc.

3. The null hypothesis is stated — “there is no difference between the population means of the two groups.”

4. A probability value for determining whether or not to reject the null hypothesis is established, e.g., if $P < .05$, then the null hypothesis will be rejected; if $P > .05$, the null hypothesis will not be rejected.

5. The data are collected and statistically analyzed. For this design a *t*-test is conducted whereby the means of the two groups are determined. The value of *t* is computed and the corresponding *P* ascertained.

¹²The term “null hypothesis” was first used by Professor Sir Ronald A. Fisher (personal communication). He chose the term “null hypothesis” without “particular regard for its etymological justification but by analogy with a usage, formerly and perhaps still current among physicists, of speaking of a null experiment, or a null method of measurement, to refer to a case in which a proposed value is inserted experimentally in the apparatus and the value is corrected, adjusted, and finally verified, when the correct value has been found; because the set-up is such, as in the Wheatstone Bridge, that a very sensitive galvanometer shows no deflection when exactly the right value has been inserted.

“The governing consideration physically is that an instrument made for direct measurement is usually much less sensitive than one which can be made to kick one way or the other according to whether too large or too small a value has been inserted.

“Without reference to the history of this usage in physics. . . . One may put it by saying that if the hypothesis is exactly true no amount of experimentation will easily give a significant discrepancy, or, that the discrepancy is null apart from the errors of random sampling.”

6. If the means are in the direction specified by the hypothesis (if the Psychology students have a higher mean score than the English students) and if the null hypothesis is rejected, it may be concluded that the hypothesis is confirmed. If the null hypothesis is not rejected, it may be concluded that the hypothesis is not confirmed. Or, if the null hypothesis is rejected, but the means are in the direction opposite to that predicted by the hypothesis, then the hypothesis is not confirmed.

"BORDERLINE" SIGNIFICANCE

One frequently occurring problem in experimentation is that of borderline significance. An experimenter who sets a P of 0.05 as his criterion for rejecting the null hypothesis fails to reject the null hypothesis if he obtains a P of 0.30. But suppose that he obtains a P of 0.06. One might argue that, "Well, this isn't quite .05 but it is so close that I'm going to reject the null hypothesis anyway. This seems reasonable; after all, this means that a difference between groups of the size that I obtained can be expected only six times out of 100 by chance when the null hypothesis is true. Surely this is not much different than a probability of five times out of 100." To this there is only one answer: the t test is *decisive* — a P of 0.06 is *not* a P of 0.05. In this case, therefore, there is no alternative but to fail to reject the null hypothesis. If the experimenter has set up a criterion of a P of 0.06 *before* he conducted his experiment, then we would have no quarrel with him — he could, in this event, reject his null hypothesis. But since he established a criterion of a P of 0.05, he cannot modify his criterion after the data are collected, not even if he obtains a P of 0.051.

At the same time, however, we must agree with this experimenter that a P of 0.06 is an unlikely event by chance. Our advice to him is: "Yes. It looks like you *might* have something. It's a good hint for further experimentation. Conduct a new experiment and see what happens. If, in this replication, you come out with significant results, you are quite safe in rejecting the null hypothesis. But if the value of t obtained is quite far from significance in this new, independent test, then you have saved yourself from making an error."

THE METHOD OF SYSTEMATIC OBSERVATION

We have previously contrasted two types of investigations: "experiments" and "systematic observation studies." The alert reader probably noticed that the two examples used to illustrate the computation of t were experiments, but that the example on p. 114 was a systematic observation study. For in the study concerning anxiety of students of Psychology and English *no variable was produced and purposely manipulated by the investigator*. Rather,

the study concerned observations of a phenomenon that was *already present in the population*. To meet the requirements for an experiment in that example we would have had to assign subjects randomly to two groups and then decree that everyone in one group would major in Psychology and all those in the other, in English. If we had been able to do this, we could say that our independent variable was "major of the student" and that we had varied it in two ways: English and Psychology majors. Since we did not vary it in this manner, it was not purposively manipulated, and hence the study cannot be said to be an experiment. In the previous two cases, however, the requirements of experimentation *were* fulfilled, since subjects were randomly assigned to two groups and then it was determined which group would receive which experimental treatment. Hence, the independent variable in the first experiment (p. 101) was the type of RNA, varied in two ways: (1) RNA from a trained group; and (2) RNA from an untrained group. Since the independent variable was under the control of the experimenter and since he *induced* different conditions in the two groups, it may be said that he *purposively manipulated* the independent variable and thus conducted an experiment.

Judgment of the importance of the difference between the two types of investigations should be held in abeyance until we consider "control" (see Chapter Six). It will be shown that there can be important differences in the two types of investigations as far as confidence in their respective conclusions is concerned. The main point to observe here is that the *t*-test may be an appropriate method of statistical analysis for both types of investigations.

SUMMARY OF THE COMPUTATION OF FOR A TWO-RANDOMIZED-GROUPS DESIGN

Assume that we have obtained the following dependent variable scores for the two groups of subjects:

Group 1	Group 2
10	8
11	9
11	12
12	12
15	12
16	13
16	14
17	15
	16
	17

1. We start with Equation (5.2), the equation for computing *t*:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{SS_1 + SS_2}{(n_1 - 1) + (n_2 - 1)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

2. Compute the sum of X (i.e., ΣX), the sum of X^2 (i.e., ΣX^2), and n for each group.

<i>Group 1</i>	<i>Group 2</i>
$\Sigma X = 108$	$\Sigma X = 128$
$\Sigma X^2 = 1512$	$\Sigma X^2 = 1712$
$n = 8$	$n = 10$

3. Using Equation (5.1), compute the means for each group.

$$\bar{X}_1 = \frac{108}{8} = 13.50 \quad \bar{X}_2 = \frac{128}{10} = 12.80$$

4. Using Equation (5.3), compute the sums of squares for each group.

$$SS_1 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1} = 1512 - \frac{(108)^2}{8} = 54.000$$

$$SS_2 = 1712 - \frac{(128)^2}{10} = 73.600$$

5. Substitute the above values in Equation (5.2).

$$t = \frac{13.50 - 12.80}{\sqrt{\left(\frac{54.000 + 73.600}{(8-1) + (10-1)}\right)\left(\frac{1}{8} + \frac{1}{10}\right)}}$$

6. Perform the operations as indicated and determine that the value of t is:

$$t = \frac{0.70}{\sqrt{(7.975)(.2250)}} = \frac{0.70}{\sqrt{1.7944}} = \frac{0.70}{1.3395} = .523$$

7. Determine the number of degrees of freedom associated with the above value of t .

$$df = N - 2 = 18 - 2 = 16$$

8. Enter the table of t , and determine the probability associated with this value of t . In this example $0.70 > P > 0.60$. Therefore, assuming a required significance level of 0.05, the null hypothesis is not rejected.

PROBLEMS (Answers are on page 376.)

1. An experimenter runs a well-designed experiment wherein $n_1 = 16$ and $n_2 = 12$. He obtains a t of 2.14. Assuming that he has set a significance level of 0.05, can he reject his null hypothesis?

2. An experimenter obtains a computed t of 2.20 with 30 df . The means of his two groups are in the direction indicated by his empirical hypothesis. Assuming that the experiment was well designed and that the experimenter

has set a significance level of 0.05, did his independent variable influence his dependent variable?

3. It is advertised that a certain tranquilizer has a curative effect on psychotics. A clinical psychologist seeks to determine whether or not this is the case. He conducts a well-designed experiment and obtains the following results on a measure of psychotic tendencies. Assuming that he has set a significance level of 0.01, and assuming that the lower the score the greater the psychotic tendency, determine whether the tranquilizer has the advertised effect.

*Scores of group
that received the
tranquilizer*

2, 3, 5, 7, 7, 8, 8, 8

*Scores of group
that did not
receive the
tranquilizer*

1, 1, 1, 2, 2, 3, 3

4. A psychologist hypothesizes that people who are of similar body build work better together. Accordingly, he forms two groups. Group 1 is composed of individuals who are of similar body build, and Group 2 consists of individuals with a different body build. He has both groups perform a task that requires a high degree of cooperation. The performance of each subject is measured where the higher the score, the better the performance on the task. He sets a significance level of 0.02. Did he confirm or disconfirm his empirical hypothesis?

Group 1

10, 12, 13, 13, 15, 15, 15, 17, 18
22, 24, 25, 25, 25, 27, 28, 30, 30

Group 2

8, 9, 9, 11, 15, 16, 16, 16, 19, 20, 21
25, 25, 26, 28, 29, 30, 30, 32, 33, 33

5. On the basis of his experience, a marriage counselor suspects that when one spouse is from the North and the other is from the South the marriage has a likelihood of being unsuccessful. He selects two groups of subjects: Group 1 composed of marriage partners both of whom are from the same section of the country (either North or South), and Group 2 consisting of marriage partners from the North and the South respectively. He sets a 0.05 level of significance and obtains ratings of the success of the marriage (the higher the rating, the better the marriage). Assume that adequate controls have been effected. Is his suspicion confirmed?

Group 1

1, 1, 1, 2, 2, 3, 3, 4, 4, 5, 6, 6, 7, 7

Group 2

1, 1, 2, 3, 4, 4, 5, 5, 6, 7

EXPERIMENTAL CONTROL

THE NATURE OF EXPERIMENTAL CONTROL

The strength of civilization is based, at rock bottom, on the amount and kinds of reliable knowledge that have been accumulated. But the progress of civilization has been slow and painfully achieved; we have, in a sense, taken two steps backward for each three steps forward. Histories of western civilization typically emphasize the backward steps, as for instance in their accounts of the great wars. At least as fascinating, though, is the record of man's achievements — the stories of the acquisition of knowledge and of the development of sound methods for acquiring that knowledge. Among the most striking advances in methodology was the recognition of the necessity for control conditions — so called “normal” conditions against which to evaluate experimental treatments. Unfortunately, cultural lag being what it is, the importance of control conditions has not yet pervaded the “logic” of everyday thinking; common-sense reasoning is most often wrong because it is based on observations of only one group (and frequently with an n of one, at that).

In order to reach the relatively advanced stage in which a proper appreciation of control conditions was recognized, we can suppose that methodologists had to first engage in considerable trial and error. There were, no doubt, a number of improperly controlled experiments conducted before methodologists became more sophisticated. And one must admire even these "semi-experiments," for they were imaginative indeed. An example is brought to our attention by Jones which "... is Herodotus' quaint account of the experiment in linguistics made by Psammetichos, King of Egypt (*Historiae* II, 2). To determine which language was the oldest, Psammetichos arranged to have two infants brought up without hearing human speech and to have their first utterances recorded. When a clear record of the children's speech had been obtained, ambassadors were sent around the world to find out where this language was spoken (specifically, where the word for "bread" was *bekos*). As a result of his experiment, Psammetichos pronounced Phrygian to be the oldest language, though he had assumed it was Egyptian before making the test" (1964, p. 419).

An account of a more sophisticated, but still ancient, investigation *did* include a control condition: "Athenaeus, in his *Feasting Philosophers* (*Deipnosophistae*, III, 84-85), describes how it was discovered that citron was an antidote for poison. It seems that a magistrate in Egypt had sentenced a group of convicted criminals to be executed by exposing them to poisonous snakes in the theater. It was reported back to him that, though the sentence had been duly carried out and all the criminals were bitten, none of them had died. The magistrate at once commenced an inquiry. He learned that when the criminals were being conducted into the theater, a market woman out of pity had given them some citron to eat. The next day, on the hypothesis that it was the citron that had saved them, the magistrate had the group divided into pairs and ordered citron fed to one of a pair but not to the other. When the two were exposed to the snakes a second time, the one who had eaten the citron suffered no harm, the other died instantly. The experiment was repeated many times and in this way (says Athenaeus) the efficacy of citron as an antidote for poison was firmly established" (Jones, 1964, p. 419).

We have emphasized, by our repeated references to the topic of control, that it is one of the most important phases in the planning and conduct of experiments. The problem of controlling variables, therefore, requires particular vigilance on the part of the experimenter. The word "control" implies that the experimenter has a certain power over the conditions of his experiment; he is able to manipulate variables in an effort to arrive at a sound conclusion. Let us illustrate by using the pharmacological example.

First, the magistrate exercised control over his independent variable by producing the event that he wished to study. This is the first sense in which we shall use the word "control." We shall say that an experimenter exercises *independent variable control* when he varies the independent variable in a known

and specified manner. In this example, the independent variable was amount of citron administered, and it was purposively varied in two ways: zero and some positive amount. (Recall from p. 59 that independent variable control is the essential defining feature of an *experiment*, as distinguished from the *method of systematic observation*).

The second sense in which we shall use "control" may be made by restating the purpose of the experiment: the magistrate sought to determine whether variation of amount of citron administered to men who were poisoned would affect their impending state of inanimation (certainly a 'clear-cut dependent variable measure, if ever there was one). He was interested in finding out whether these two variables were related. There were also present, however, a number of other (extraneous) variables that might have affected the subjects' degree of viability. If there was, in fact, a relationship of the type that he sought, it might have been hidden from him by these other variables. Some substance in the subjects' breakfast, for instance, might have been an antidote; the subjects might have been members of a snake cult and thereby developed an immunity; and so forth. In the absence of knowledge of such extraneous variables, it was necessary to assume that they might have affected the dependent variable. Hence, their possible effects were controlled, i.e., the magistrate formed two equivalent groups and administered citron to only one. In this way, the two groups were equated with regard to all extraneous variables so that their only difference was that one received the hypothesized antidote. The fact that only members of the group that received citron survived ruled out further consideration of the extraneous variables. With this control effected, the magistrate obtained the relationship that he sought, and our second sense of "control" is illustrated: *Extraneous variable control* refers to the regulation of extraneous variables.

In order to be clear we shall say that an extraneous variable is one that is operating in the experimental situation in addition to the independent variable. Since the extraneous variable might affect the dependent variable, and since we are not immediately interested in ascertaining whether or not it does affect the dependent variable, it must be regulated so that it will not mask the possible effect of the independent variable.

Failing to control extraneous variables results in a *confounded experiment*, a disastrous consequence for the experimenter, i.e., if an extraneous variable is allowed to operate in an uncontrolled manner, it and the independent variable are confounded (the dependent variable is not free from irrelevant influences). Suppose, for example, that the subjects who received citron had been served a different breakfast than the control subjects. In this case the magistrate would not know whether it was citron or something in the breakfast of the experimental subjects that was the antidote — type of breakfast would thus have been an extraneous variable that was confounded with the independent variable. We can thus see that confounding occurs when there

is an extraneous variable that is systematically related to the independent variable, and it *may* act on the dependent variable; hence, the extraneous variable may affect the dependent variable scores of one group, but not the other. If confounding is present, then, the reason that any change occurs in the dependent variable cannot be ascribed to the independent variable. In summary, *confounding occurs when an extraneous variable is systematically related to the independent variable, and it might differentially affect the dependent variable scores of the two groups.*

To illustrate further these two senses of "control," and in particular to get closer to home, consider a psychological example. Suppose that an experimenter is interested in determining the effect of Vitamin A on certain visual abilities. The independent variable might be operationally defined as the amount of Vitamin A administered according to a certain schedule. The dependent variable might be similarly defined as the number of letters that a subject can see on a chart placed some distance from him. Since the independent variable is under the control of the experimenter he may vary it as he wishes. He may, for instance, vary it in three ways: one group of subjects may receive a placebo but no Vitamin A; a second group may receive a total of three units of the vitamin, while a third group is administered a total of five units. In this way he is exercising control of the independent variable.

To illustrate extraneous variable control we might note that the lighting conditions under which the test is taken are relevant to the number of letters that the subjects can correctly report. Suppose, for example, that the vision test is taken in a room in which the amount of light varies considerably during the day, and further that Group 1 is run mainly in the morning, Group 2 around noon, and Group 3 in the afternoon. In this case some subjects would take the test when there is good light, others when it is poor. The test scores might then primarily reflect the lighting conditions rather than the amount of Vitamin A administered, in which case the possible effects of Vitamin A would be masked out. Put another way, the amount of lighting and amount of Vitamin A would be confounded. Lack of control over this extraneous variable would leave us in a situation where we do not know which variable or combination of variables is responsible for influencing our dependent variable.

Just to develop this point briefly, let us consider some of the possibilities when only the single extraneous variable of light is uncontrolled. Assume that the obtained value of the dependent variable increases as the amount of Vitamin A increases, i.e., that the group receiving the five-unit dose of Vitamin A has the highest dependent variable score, the three-unit group is next, and that the zero Vitamin A group has the lowest test score. What may we conclude about the effect of Vitamin A on the dependent variable? Since light is uncontrolled we do not know what effect it has. Hence, the light may actually be the factor that causes the dependent variable scores to

increase. Or, it is possible that lighting has a detrimental effect such that if it were not operating in an uncontrolled fashion the apparent effects of Vitamin A would be even more pronounced, e.g., if the five-unit group received a score of ten, it might have received a score of 20, if light had been controlled. Another possibility is that the light has no effect, in which case our results could be accepted as valid. But since we do not know this, we cannot reach such a conclusion. The ambiguity in interpreting the effects of an independent variable where a single extraneous variable is not controlled should thus be apparent. But where there is more than one extraneous variable that is uncontrolled, the situation is much nearer total chaos.

Experimental control, then, is the regulation of experimental variables. And we may consider two classes of experimental variables: independent and extraneous. The independent variables, we have said, are those whose effects the experimenter is attempting to determine. He wants to know if a given independent variable affects his dependent variable. The extraneous variables are all other variables operating on the subjects at the time of the experiment. By exercising independent variable control the experimenter varies the independent variable as he wishes. By exercising extraneous variable control he regulates the extraneous variables so that confounding is eliminated. If adequate extraneous variable control is exercised, an unambiguous statement on the relationship between the independent and dependent variables can be made. If extraneous variable control is inadequate, however, the conclusion must be tempered. The extent to which it must be tempered depends on a number of factors, but, generally, inadequate extraneous variable control leads to no conclusion whatsoever concerning the relationship.

DETERMINING EXTRANEOUS VARIABLES

We know that at any given moment a fantastically large number of stimuli are impinging on an organism. And we must assume that all of these stimuli are affecting the organism's behavior. But in any given experiment we are usually interested in only one aspect of behavior — a single class of responses. Furthermore, we usually seek to determine whether a certain class of stimuli affect that response; this is the independent-dependent variable relationship. Hence, for this immediate purpose we want to eliminate from consideration all other variables. If this were possible we could conclude that any change in our dependent variable is due only to the variation of our independent variable.

If these other (extraneous) variables are allowed to influence our dependent variable, however, any change in our dependent variable could not be ascribed to variation of our independent variable. *We would not know which of the numerous variables caused the change.*

We must, then, control the experimental situation so that these other, extraneous variables can be dismissed from further consideration. The first step in this process is to identify them: what extraneous variables may be present in the experimental situation? Since it would be an almost endless task to list all of the variables that *might* affect the behavior of an organism, our question must be more limited: Of all the variables present, which might *conceivably* affect our dependent variable? Even though this is still a difficult question, we can immediately eliminate from consideration a large number of unlikely influences on the organism. For example, if we are studying a learning process, we would not even consider such variables as color of the chair in which the subject sits, brand of pencil he uses, etc. As a first step in determining those extraneous variables that should be considered, we might refer to our literature survey. We can study previous experiments concerned with our dependent variable to find out which variables have been demonstrated to affect that dependent variable. We should also note what other variables previous experimenters have considered it necessary to control. Discussion sections of earlier articles may also yield information about variables that had not previously been controlled, but were recommended for consideration in the future. Together with the results of our literature survey, our general knowledge of potentially relevant variables, and considerable reflection concerning other variables, we may arrive at a list of extraneous variables that should be considered.

SPECIFYING EXTRANEOUS VARIABLES TO BE CONTROLLED

Once our list of potentially relevant extraneous variables is constructed, we must decide which should be controlled. This would include those variables that are likely to affect our dependent variable. It is to these variables that the various techniques of control will be applied. A discussion of these techniques is presented on pp. 127-137. Suffice it to state here the end result — the changes in the dependent variable will be ascribable to the independent variable rather than to the controlled extraneous variables.

SPECIFYING EXTRANEOUS VARIABLES THAT CANNOT REASONABLY BE CONTROLLED

A simple answer to the question of which extraneous variables should be controlled is that we should control all of them. Although it *might* be possible to control all of them, such a feat would be too expensive in terms of time, effort, and money. For example, suppose that the variation in temperature during experimental sessions is five degrees. Even though it is possible to control this variable, it is highly unlikely (in most experiments) that it would significantly (if at all) affect the dependent variable. And the experimenter's

effort would be great enough to make the game "not worth the candle." This is particularly so when one considers the large number of other variables in the same category. With the limited amount of energy and resources available to him, the experimenter should seek to control only those variables that he considers potentially relevant.

Now, all of these probably minor variables might accumulate to have a rather major effect on the dependent variable, thus invalidating the experiment. And even if the effect is not so extreme, should even a minor extraneous variable be allowed to influence the dependent variable? If the experimenter is not going to control them, what can he do about them? In thinking about these points, we must remember that there always will be a large number of variables in this category. The question is, will they affect one of our experimental conditions (one of our groups) to a greater extent than another? For, if by chance such variables do not differentially affect our groups, then our worries are considerably lessened. We can make the assumption that such variables will "randomize out," that, in the long run, they will affect both groups equally. If it is reasonable to make this assumption, then this type of variable should not delay us further. A further discussion of randomization as a technique of control will allow us to consider this and similar problems later.

WHEN TO ABANDON THE EXPERIMENT

Up to this point we have been rather optimistic. We have assumed that we are capable of controlling all of the relevant variables that affect the dependent variable, that the effects of these variables will be essentially equal on all groups in the experiment. If it is unreasonable to make this assumption, then the experimenter must ask himself if these variables are of sufficient importance to necessitate the abandonment of the experiment. Even if he is not sure on this point, perhaps it would be best if the experiment were not conducted. Sometime after considering the various control problems, the experimenter must ask himself what he will gain by conducting the experiment. In these cases of inadequate extraneous variable control, the answer need not be that nothing will be gained; for instance, by conducting such an experiment it may be that further insight or beneficial information will be acquired concerning the control problem. But the experimenter must realize that this situation exists and be realistic in understanding that it may be better to discontinue the project.

TECHNIQUES OF CONTROL

We have previously emphasized the importance of exercising adequate experimental control, but this phase of the experimental planning is sufficiently important that it does not seem possible to *overemphasize* it. Although

experimenters try to exercise considerable vigilance in this regard, it is frequently the case that a crucial, uncontrolled extraneous variable is discovered only after the data are compiled. Shortcomings in control are found even in published experiments. Certainly, if such variables have been discovered neither by the experimenter nor by the editors of the journals, they are elusive and subtle. Furthermore, errors of control are not the sole property of young experimenters; they may be found in the work of some of the most respected and established psychologists.

The experimenter should give as much thought to potential errors as possible. After he has checked and rechecked himself it may be possible to obtain critiques from colleagues. An "outsider" can sometimes approach the experiment with a totally different set, thus seeing something that the experimenter himself might have missed.

Our main consideration in this section follows from the point at which an important extraneous variable is spotted and the experimenter must ask himself how it is to be controlled. He must ascertain what techniques are available for regulating it in such a manner that the effects of the independent variable on the dependent variable can be clearly isolated. There are a number of such techniques; we shall attempt to classify them into several categories. This classification will necessarily be incomplete and overlapping in part, particularly as to the variations of each class. But a general understanding of the major principles should facilitate their application to a wide variety of specific control problems.

1. *Elimination.* The most desirable way to control extraneous variables is to eliminate them from the experimental situation. Examples of the elimination of extraneous variables in psychological laboratories would be the use of sound-deadened rooms or the Skinner box, which is sound-deadened and opaque. Unfortunately, though, most extraneous variables cannot be eliminated. The previous example concerning Vitamin A and the subject's ability to read letters from a chart is a case in point. In that example our extraneous variable was the amount of lighting. Obviously, if the method of elimination were applied, the subjects would not have the light needed for them to see the chart. Other extraneous variables that one would have a hard time eliminating are subjects' previous experience, sex, level of motivation, age, weight, intelligence, and so on.

2. *Constancy of Conditions.* When certain extraneous variables cannot be eliminated, we can attempt to hold them constant throughout the experiment. Control by this technique means essentially that whatever the extraneous variable, the same value of it is present for all subjects. Perhaps, for instance, the time of day is an important variable. Maybe people perform better on the dependent variable early in the morning than late in the afternoon. In order to hold time of day constant, we might introduce all subjects into the experimental situation at approximately the same hour on successive days.

Of course this procedure would not really hold the amount of fatigue constant for all subjects on all days. But it would certainly help.

Another example of effecting constancy of conditions would be our Vitamin A chart-reading experiment. In this case we would attempt to hold the lighting conditions constant. Thus, we might pull down the blinds in our experimental room and have the same light turned on for all subjects. In experiments where light intensity is extremely important, we could actually measure the amount of light present for each subject. The placing of a rheostat in the lighting circuit would allow us to modify fluctuations in the electrical flow in such a manner as to hold light intensity at almost precisely the same value for all subjects. Or we might prefer to use a DC source of electricity for our light as it would not fluctuate.

One of the standard applications of the technique of holding conditions constant is to conduct experimental sessions in the same room. Thus whatever might be the influence of the particular characteristics of the room (gayness, odors, color of the walls and furniture, location), that influence would be the same for all subjects. In like manner to hold various subject variables constant (educational level, sex, age), we need merely select subjects with the characteristics that we want. For example we might specify that all subjects must have completed the eighth grade and no more; that all subjects are male; or that all subjects are 50 years old.

Numerous characteristics of our experimental procedure must be subjected to this technique of control. Instructions to subjects, for instance, are extremely important. For this reason experimenters read precisely the same written set of instructions to all subjects (except where they must be modified for different experimental conditions). But even if the same words are read to all subjects, they might be read in different ways, with different intonations and emphases, regardless of the experimenter's efforts to avoid such differences. To exercise more precise control, then, many experimenters have subjects listen to the same standardized instructions from a tape recorder.

Procedurally, all subjects should go through the same steps in the same order. For instance, if the steps are: greet subject, seat him, read instructions, blindfold him, tell him to start, and so on, then one would not want to blindfold some subjects *before* the instructions were read and blindfold others *after* the instructions. The attitude of the experimenter should also be held as constant as possible for all subjects. If he is jovial with one subject, and gruff with another, confounding of experimenter attitude with the independent variable would occur. Now, acting the same toward all subjects is extremely difficult, but a strong effort should be made in this direction. The experimenter can practice the experimental procedure a number of times until it becomes so routine that he can treat each subject in a mechanical fashion. Of course, the same experimenter should collect data from all the subjects. If different experimenters are used unsystematically, then a rather

serious error may result. In one experiment, for instance, an experimenter ran a group of rats for 14 days, but had to be absent on the 15th day. The rats' performance for a different experimenter on that day was sufficiently different from other groups who had not suffered a change of experimenters that it is reasonable to conclude that the mere handling of them by a new person (who undoubtedly used somewhat different methods of picking them up, etc.) was responsible for the change.

The apparatus that is used both in administering the experimental treatment and in recording the results should be the same for all subjects. Suppose, for example, that two memory drums are used in an experiment, one of which moves more slowly than the other. If one group uses the faster drum and another the slower, confounding will result. Application of the technique of constancy of conditions dictates that all subjects use the same drum. Similar precautions should be taken with regard to recording apparatus.

3. *Balancing.* When it is not convenient or possible to hold constant conditions in the experiment, the experimenter may resort to the technique of balancing out the effect of extraneous variables. There are two general situations in which balancing may be used: (1) where the experimenter is either unable or uninterested in identifying the extraneous variables; (2) where he can identify them and desires to take special steps to control them.

Consider the first situation. One group of experimenters were interested in the effect of rifle training on rifle steadiness; whether a prolonged period of training in rifle firing increased the steadiness with which soldiers held their weapons (McGuigan & MacCaslin, 1955b). Previous research had indicated that the steadier a man held a rifle, the more accurately he could shoot. Thus, if you could increase steadiness through rifle training, you *might* thereby increase rifle accuracy. The design of the experiment was a test of rifle steadiness before and after subjects received their rifle training. If the soldiers were steadier on the second test, it might be concluded that training increases steadiness. The first group of data that were analyzed are presented in Table 6.1, where the lower the score, the greater the steadiness.

Table 6.1. *Mean Steadiness Scores of Soldiers Before and After Rifle Training.*

<i>Before Training</i>	<i>Training Period</i>	<i>After Training</i>
235.39		194.26

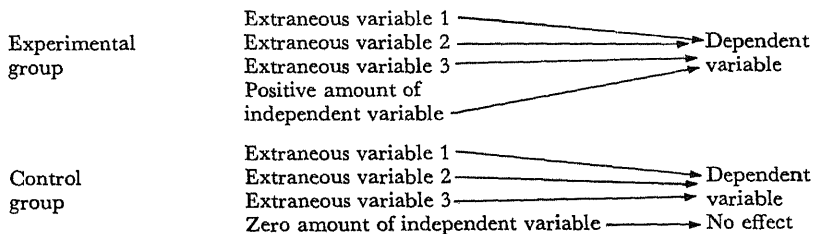
From Table 6.1 we can see that the scores actually did decrease. The first thought is to conclude that training increases steadiness. When the experimenters analyzed another set of data from a control group which did not receive rifle training, the picture changes (see Table 6.2).

Table 6.2. *Mean Steadiness Scores of Trained and Untrained Groups.*

	Before Training		After Training
Trained (Experimental) Groups	235.39	Training period	194.26
Untrained (Control) Group	226.61	No training period	170.33

From Table 6.2 we can see that not only did the steadiness scores of the untrained group also decrease, but that they decreased more than those of the trained group. In order to reach the conclusion that training is the variable responsible for the decrease in scores, the experimental group had to show a significantly greater decrease in scores than did the control (no training) group. Thus, we may say that rifle training was not the reason that the steadiness scores of the trained group decreased. There must have been other variables operating to produce that change, variables that operated on both the experimental and the control groups. Whatever the variables, they were controlled by the technique of balancing (i.e., their effects on the trained group were balanced out or equalized by the use of the control group). But we may speculate about these extraneous variables. For example, the rifle training was given during the first two weeks of the soldiers' army life. It may be that the drop in scores merely reflected a general adjustment to the emotional impacts of army life. Or the soldiers could have learned enough about the steadiness test in the first session, to improve their performance in the second.

But whatever the extraneous variables, they were controlled by the use of the control group. The logic of using a control group should now be apparent. If the experimental and control groups are treated in the same way except with regard to the independent variable, then any difference between the two groups on the dependent variable is ascribable to the independent variable (at least in the long run). Thus, we need not specify all the extraneous variables that influence the two groups during the experiment. For instance, suppose that only three extraneous variables operate on the experimental

**FIGURE 6.1.**

Diagrammatic representation of the use of the control group as a technique of balancing.

group in addition to some positive amount of the independent variable. By administering a zero amount of the independent variable to the control group and by balancing out the effects of the three extraneous variables by allowing them to operate also on the control group, we can see from Figure 6.1 that the independent variable is the only one that can differentially influence the two groups.

An additional important use of the control group as a technique of control may now be profitably discussed. Granting that: (1) a large number of extraneous variables are operating on a subject in any given situation; and (2) we cannot remove all of these variables, *then* we can use additional control groups to evaluate the influence of these variables, to analyze the total situation into its parts. Referring to Figure 6.1 we may be interested in the effect of extraneous variable 1. To evaluate that extraneous variable we need

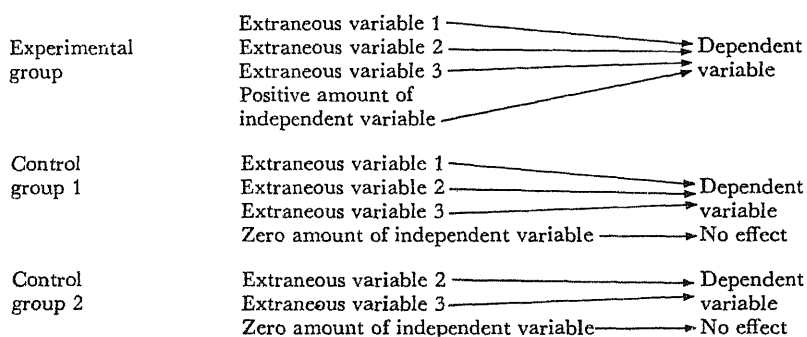


FIGURE 6.2.

The use of a second control group to evaluate the effect of an extraneous variable.

only add an additional control group which is not influenced by it (i.e., receives a zero value of it). The plan is illustrated in Figure 6.2.

For both the experimental group and control group 1, extraneous variable 1 is possibly influencing the dependent variable. Since this variable is not operating for control group 2, a comparison of the two control groups should tell us the effect of extraneous variable 1. Consider one of the extraneous variables that was operating in the rifle steadiness experiment: practice in the test situation. If an acquaintance with testing procedure and the specific learning of how to hold the rifle while tested led to lower scores, the addition of control group 2 should provide us with this information (see Table 6.3).

This is the same design as in Table 6.2 except for the addition of a second control group that does not take the initial test. A comparison of the steadiness scores of the two untrained groups on the test after training should tell us what the effect of the initial test is. If, for instance, Untrained Group 2 is less

Table 6.3. *Possible Experimental Design for Studying the Effect of Practice on the First Steadiness Test.*

	<i>Receive Test Before Training?</i>	<i>Receive Training?</i>	<i>Receive Test After Training?</i>
Trained (Experimental) Group	yes	yes	yes
Untrained (Control) Group 1	yes	no	yes
Untrained (Control) Group 2	no	no	yes

steady on the second (post-training) test than Untrained Group 1, we could say that the effect of taking the first (pretraining) test is to increase steadiness.

The use of additional control groups to evaluate the effect of various extraneous variables is a very important experimental technique. Frequently, the independent variable itself is of such a nature that it can be further fractionated by the addition of control groups. Some additional examples of this technique are taken up in the last section of this chapter. For additional study of this topic see the excellent article by Solomon (1949).

The second situation in which balancing may be used is where there is a specific and known extraneous variable to be controlled. For instance, if an experimenter wishes to control the sex variable, he can use subjects of only one gender. If he has both male and female subjects available, however, and not enough subjects of one sex to use them exclusively, he may be forced to use both sexes. In this event he may control the effect of gender on his dependent variable by making sure that it balances out in his two groups. This would be accomplished by assigning an equal number of subjects of each sex to each group. If he has 40 males and 30 females, he would make sure that 20 males and 15 females go in each group.¹ Thus, if sex is relevant to his dependent variable, its effects would be the same for each experimental condition. In a similar manner he could control the age of the subjects; he would make sure that an equal number of each age classification is assigned to each group.

The same holds true for apparatus problems. Suppose that an experimenter has two memory drums available and wants to use both to save time. They may or may not have slight differences, but to make sure that this variable is controlled he would have half of the subjects in each group use each memory drum. If he has 30 subjects in each of two groups, 15 subjects in each group would use drum A, while the other 15 would use drum B. Thus, whatever the differences in the memory drums, if any, their respective effects would be the same for both groups.

¹Of course, in this example, even though the effect of sex is balanced out, the males will have a greater effect on the dependent variable scores than the females. The matter can be handled with appropriate statistical procedures, but it is preferable to have an equal number of each gender assigned to each group, in this example.

Balancing may also be applied where there is more than one experimenter. In this case we merely need to have each experimenter run an equal number of the subjects in each group. To consider a situation that is a bit more complicated than any we have previously discussed let us say that we wish to balance two effects: sex and experimenter. We have two groups: 60 subjects per group, two sexes, and two experimenters. In this case the balancing arrangement would look something like that presented in Table 6.4.

In a final example of the balancing technique, let us say that we want to see how well rats retain a certain habit. To do this we might use a T maze which has a white and a black goal box. From the start position, the rats may run to either box. But we train them to run to one box, say the black. Hence, we always feed them when they run to the black box; they receive no food in the white box. After a number of trials they run rather consistently to the

Table 6.4. *Illustration of a Design where Experimenters and Sex are Balanced.*

<i>Group I</i>	<i>Group II</i>
15 Males — Experimenter 1	15 Males — Experimenter 1
15 Males — Experimenter 2	15 Males — Experimenter 2
15 Females — Experimenter 1	15 Females — Experimenter 1
15 Females — Experimenter 2	15 Females — Experimenter 2

black box, avoiding the white box. After this initial training, we do not allow them to run the maze for, say, three months, at the end of which we place them in the start box of the maze for their test trials. If most of their runs are to the black box, may we assume that they “remembered” well? Our conclusion should not be so hasty. For we know that rats are nocturnal animals and that they tend to prefer dark places over light places. In particular, they have a preference for black goal boxes over white ones *before any training*. Hence it is possible that they would go more frequently to the black box on the test trials *regardless of the previous training*, and thus may not have “remembered” anything. For this reason we need to exercise control over the color of the “reward” box — we need to balance the colors. To do this we train half of our animals as above. The other half are trained in the opposite manner; they receive food when they run to the white box but no food when they run to the black box. If, on the test trials, the animals trained to go to white show a preference for white, then we would have considerably more confidence in our conclusion. For regardless of the color that they were trained to, we find that they retain the habit over the three month period. The effect of color cannot possibly be the variable that is influencing their behavior.

4. *Counterbalancing.* Some experiments are designed so that the same subject must serve under two or more different experimental conditions. If

an experimenter were interested in whether a stop sign should be painted yellow or red, his problem would be to determine to which colored sign a subject responds faster. To answer this question he might measure a subject's reaction time to, first the yellow sign, and then the red sign. By repeating this procedure with a number of subjects he could reach a conclusion, perhaps that reaction time to the red sign is the smaller. Since the subjects were first exposed to the yellow sign, however, their reaction time to that sign would be partially dependent on their learning to operate the experimental apparatus and on their adaptation to the experimental situation. After they have learned how to operate the apparatus and adapted to the situation, they are exposed to the red sign. Hence, their lower reaction time to the red might merely reflect practice and adaptation effects rather than effect of color — color of sign and amount of practice are confounded. The answer frequently given to the problem of how to control the extraneous variable of amount of practice is to use the method of counterbalancing. The application of this method would be to have half the subjects react to the yellow sign first and the red sign second, while the other half would experience the red sign first and the yellow sign second (see Table 6.5).

Table 6.5. *Demonstrating Counterbalancing to Control an Extraneous Variable.*

	EXPERIMENTAL SESSION	
	1	2
$\frac{1}{2}$ Subjects	Yellow Sign	Red Sign
$\frac{1}{2}$ Subjects	Red Sign	Yellow Sign

The general principle of the technique of counterbalancing may be stated as: Each condition (e.g., color of sign) must be presented to each subject an equal number of times, and each condition must occur an equal number of times at each practice session. Furthermore, each condition must precede and follow all other conditions an equal number of times. It can be seen that this principle is applicable to any number of conditions. For example, the principle of counterbalancing could be applied where we have three colors of signs, in which case one-sixth of the subjects would be presented with each order specified in Table 6.6. Color of sign presented at each session is indicated by R (Red), Y (Yellow), or G (Green).

By studying Table 6.6, we can observe that the requirements for counterbalancing the effects of three variables are satisfied. In particular, if six subjects are used (any multiple of six such as 12 or 18 would suffice for this design), we can observe that each color of sign is presented twice at each session, that each subject receives each color once, and that each color precedes and follows each other color two times.

Table 6.6. *A Counterbalanced Design For Three Independent Variables.*

	EXPERIMENTAL SESSION		
	1	2	3
1/6 Subjects	R	Y	G
1/6 Subjects	R	G	Y
1/6 Subjects	Y	R	G
1/6 Subjects	Y	G	R
1/6 Subjects	G	R	Y
1/6 Subjects	G	Y	R

If a number of experimental sessions are involved, not only would a certain amount of improvement in the subjects' performance due to practice be expected, but also a certain decrement in performance due to fatigue. The method of counterbalancing attempts to distribute these effects equally to all conditions. Hence, whatever the practice and fatigue effects, they presumably influence each condition equally since each condition occurs equally often at each stage of practice.

In extending the general principle of counterbalancing to a large number of variables, the number of orders (and therefore the numbers of subjects required) soon becomes unrealistic. For example, to counterbalance four variables one would require at least 24 subjects; for five variables, there are 120 orders; and for six variables, a minimum of 720 subjects would be run. To solve this problem, one may resort to *incomplete* (as against *complete*) counterbalancing. An incomplete counterbalancing design still requires that each subject receive each treatment once, and only once, and that each treatment occur an equal number of times (once) during each session; but it does not require all possible orders of the variables to be presented. Incomplete counterbalancing still adequately controls practice or fatigue effects, and it is generally recommended by Underwood (1966). Another presentation of designs that may be used in incomplete counterbalancing is given in Edwards (1960, pp. 275-76).

For a more extensive discussion of this topic, read the excellent presentation by Underwood (1966, pp. 459-66). Let us here complete our discussion by pointing out a difficulty that one might encounter in counterbalancing: In using the technique, one assumes that the effect of presenting one variable before a second is the same as presenting the second before the first, e.g., that the practice effects of responding to the red sign first are the same as for responding to the yellow sign first. This might not be the case, so that seeing the red sign first might induce a greater practice (or fatigue) effect, possibly leading to erroneous conclusions. More generally, counterbalanced designs entail the assumption that there is no differential (asymmetrical) transfer between conditions. By *differential* or *asymmetrical transfer* we mean that the transfer from condition one (when it occurs first) to condition two is different than the transfer from condition two (when it occurs first) to condition one.

If this assumption is not justified, there will be interactions (see p. 249) among the (order and treatment) variables that will lead to difficulties in the statistical analysis (cf. Gaito, 1958). That differential transfer in counterbalanced designs is not a completely unlikely possibility has been pointed out by Poulton and Freeman (1966). In one example, the purpose was to study the effects of variation of air pressure on card sorting behavior. One group of men first sorted cards when the pressure surrounding them was 3.5 atmospheres absolute (call this condition A); they then sorted cards at a normal pressure (condition B). A second group of men experienced condition B first, followed by condition A. Many slow responses occurred for the first group of men under condition A, as one might expect. But when they sorted cards under normal pressure, these men made almost as many slow responses as they did under condition A. The second group, on the other hand, made a fewer number of slow responses under normal pressure (condition B), and made almost the same number of slow responses when they shifted to condition A. In other words, card sorting behavior (the dependent variable) was influenced by the *order* of presenting the experimental conditions. As a result, when the results for the first and second sessions were combined, the effect of variation of pressure was obscured and the statistical test indicated (erroneously) that it was not a reliable effect. In general, these authors conclude, asymmetrical transfer reduces, but can exaggerate, the difference between two conditions.

The lesson, then, is that if you use counterbalancing as a technique of control, you should examine your data for asymmetrical transfer effects. If you find yourself in possession of such, you might study your findings as interactions (p. 249-253) and consult the more detailed advice given by Poulton and Freeman (1966). To end on a happy note, however, let us observe that Underwood (1966) emphasizes the many advantages of counterbalancing and recommends its use "for many experiments."

To eliminate confusion let us specify a major difference between balancing and counterbalancing. Balancing is used when each subject is exposed to only one experimental condition. Counterbalancing is used only when there are repeated measures on (more than one condition for) the same subject.

5. *Randomization.*² This technique is used for two general situations: (1) where it is known that certain extraneous variables operate in the experimental situation, but it is not feasible to apply one of the above techniques of control; (2) where we assume that some extraneous variables will operate, but cannot specify them and therefore cannot apply the other techniques.³ In either case we take precautions that enhance the likelihood of our assumption

²Randomization is included as a technique of control because the experimenter takes certain steps to insure its operation and thus equalizes the effects of extraneous variables.

³An additional reason that randomization is important (such as random assignment of subjects to groups) is to insure the validity of the statistical test. But this point is covered later under the topic of assumptions of statistical tests (p. 354).

that the extraneous variables will "randomize out," i.e., that whatever their effects, they will influence both groups to approximately the same extent.

Consider some examples. Subjects' characteristics are important in any psychological experiment. Such variables as previous learning experiences, level of motivation, amount of food eaten on the experimental day, relations with boy or girl friends, and money problems, may affect our dependent variable. Of course, the experimenter cannot control such variables by any of the previous techniques. If, however, he has an experimental and a control group, say, and if he has randomly assigned subjects to the two groups, he may assume that the effect of such variables is about the same on both groups. He may expect the two groups to differ on such variables only within the limits of random sampling. Hence, the extraneous variables should not differentially affect his dependent variable. And whatever the differences (small, we expect) between the groups on such variables, they are taken into account by our method of statistical analysis. For instance, the *t*-test is designed so that it will tell us whether the groups differ on other than a basis of random fluctuations.

One of the most incredible examples of confounding that the author has ever encountered occurred because the experimenter failed to randomly assign his subjects to groups. It would not have been incredible, perhaps, had it been committed by a high school student, but it happens that it occurred for a master's thesis. Without going into the details of the experiment, the student used speed of running a maze as an index of the strength of a theoretical variable. His reasoning was to the effect that Group 1 should have a larger amount of the theoretical variable present (due to various training conditions), and should thus have the greater speed. But in retrospect, the experiment never got off the ground. In assigning rats to groups, the "experimenter" merely reached into the cages and the first subjects that came into his hands were assigned to Group 1, the remainder to Group 2. The more active animals no doubt popped their heads out of the cage to be grasped by the experimenter, while the less active ones cowered in the rear of the cage. Regardless, therefore, of the training administered to the groups, Group 1 had, in all likelihood, the speedier rats. Experimental treatments and initial (native) running ability were thus confounded. This example, incidentally, serves to justify what otherwise might seem as an arbitrary decision to include randomization as a technique of control. For if the experimenter does not take specific steps to assure randomization (such as randomly assigning subjects to groups) he can become the victim of a confounded experiment.

The potential extraneous variables that might appear in the experimental situation are considerable. Various events might occur in an unsystematic way, such as the ringing of campus bells, the clanging of radiator pipes,

peculiar momentary behavior of the experimenter (such as a tic, sneezing, or scratching), an outsider intruding, odors from the chemistry laboratory, and the dripping of water from an overhead pipe. Now it might be possible to anticipate many of these variables and control them with one of our techniques, but even if it is possible, it might not be feasible. Signs may be placed on the door of the laboratory to head off intrusions, but signs are not always read. A sound-deadened room is the answer to many of the problems, but such facilities are not always available in psychological laboratories. It is unlikely that *all* such variables will be controlled by means of the previous techniques. Accordingly we can do the next best thing — we take steps to assure that their effects will randomize out so that they will not differentially affect our groups. To facilitate the credibility of this assumption we might make sure that the order in which we run our subjects is approximately that of alternation. Thus, if we randomly assign the first subject we run to the experimental group, the next would be in the control group; the third subject would be randomly assigned to either the control or experimental group, whereupon the fourth would be in the alternative group; and so forth. In this way we could expect, for example, that if a building construction operation is going on that is particularly bothersome, it will affect several subjects in each group and both groups approximately equally.

AN EXAMPLE OF EXERCISING EXTRANEOUS VARIABLE CONTROL

To illustrate some of our major points, and to try to unify our thinking about control procedures, consider an experiment that has as its purpose the determination of whether the amount of stress present on members of a group influences the amount of hostility they will verbally express toward their parents while talking in that group situation. To answer this question we would first plan on collecting a number of individuals. Since we need to vary the amount of stress present on the members, we form two groups. A fairly heightened amount of stress is exerted on the experimental group (by some means that need not detain us here), while the control group experiences only the normal stress present in such a social situation. Our independent variable is thus amount of stress (which is varied in two ways), and the dependent variable is amount of hostility verbally expressed toward parents. Referring to Figure 6.3 we note that, as far as control is concerned, our first step is to determine the extraneous variables that are present. Through the procedures previously specified we might arrive at the following list: sex and age of subjects, whether their parents are living or dead, place of the experiment, time of day, characteristics of experimenter, lighting conditions, various noises, number in the groups, family background and ethnic origin of

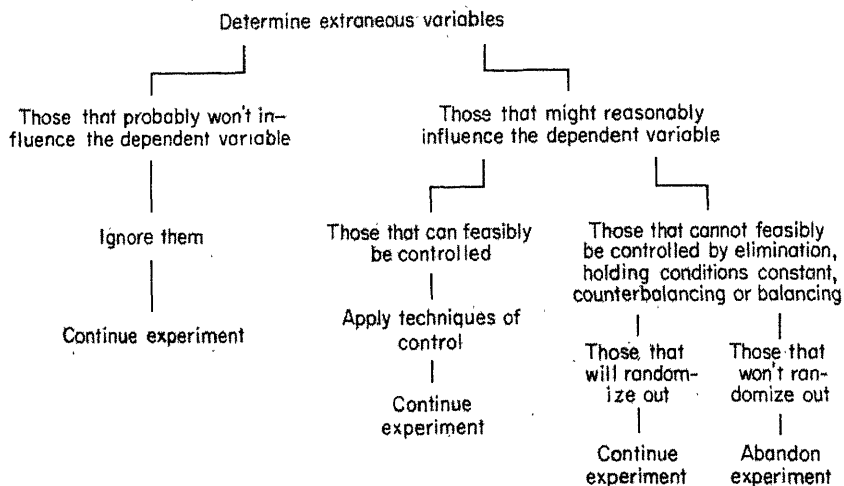


FIGURE 6.3.

An overall diagram of steps to be followed in planning an experiment.

subjects, their educational level, recent experiences with parents, general aggressive tendencies, frustrations, previous feelings towards parents, and eye color.

From Figure 6.3 we note that the next step is to determine those extraneous variables that might reasonably influence the dependent variable. Merely for illustrative purposes we have included one that probably will not influence the dependent variable and thus will be ignored: eye color of subjects.⁴ All the rest must be dealt with. Those that might feasibly be controlled by elimination, holding conditions constant, or balancing, we might decide, are sex, age, place, time, lighting, group number, education, whether parents are living or dead, and experimenter characteristics. Of these, we could control the following by holding conditions constant: place of experiment — by holding sessions for both groups in the same room; time of day — by holding sessions at the same time for both groups (on different days); lighting conditions — by having the same lights on for both groups with no external light present; number in the groups — by having the same number in each group; and experimenter characteristics — by having the same experimenter, with the same experimenter attitude, appear before both groups.

The variables of sex, age, educational level, and parents living or dead could be controlled by balancing. We could assign an equal number of each sex to each group, make sure that the average age of each group is

⁴Even though there are variables that we ignore we have several times made the point that they are, in actuality, controlled through the techniques of balancing and randomization.

about the same, distribute educational level equally between both groups, and assign an equal number of subjects whose parents are living to each group. Now, it is obvious that simultaneous balancing of all these variables would be rather difficult (if not impossible) with a small number of subjects. In fact, two variables would probably be as many as we could feasibly handle by this technique. We might select sex and parents living or dead as the most important and balance them out. For, if we are using college students (as we probably would), educational level and age would be about the same. Hence, we shall lump these two variables with the following as variables that we do not consider it feasible to control by the above techniques: various noises, family background, ethnic origin, recent experiences with parents, general aggressive tendencies, extent of frustration, and previous feelings toward parents. Some of these might be considered important, and it would certainly be desirable to control them. Most of them are difficult to measure, however, and thus are hard to balance out in a specific manner.

Now, is it reasonable to assume that such variables will randomly affect both groups to approximately the same extent? If subjects are randomly assigned to groups (except insofar as we have restricted the random process through balancing), the assumption should be valid. And as previously noted we can always check on the validity of this assumption by comparing the groups on any given variable. Since this assumption seems reasonable to make in the present example we shall conclude that the conduct of the experiment is feasible as far as control problems are concerned; we have not been able to specify any confounding effects that would suggest that the experiment should be abandoned.

THE EXPERIMENTER AS AN EXTRANEOUS VARIABLE

We have several times mentioned the matter of controlling experimenter influences on the dependent variable, but the topic is sufficiently important that we shall concentrate on it here, and in a later section, too (pp. 329-343). Even though we have long been aware that experimenter characteristics may have a substantial effect on subjects, researchers have essentially ignored this variable in the design of their experiments. To document this point, a sample of articles from the *Journal of Experimental Psychology* was studied (McGuigan, 1963). The conclusion was that although more than one person collected data in a large number of the experiments reported, in no case was any mention made of techniques of controlling the experimenter variable, and in only one of the sample of articles was the number of data collectors actually specified. The possibility is alarming that in these studies adequate control of the experimenter variable was not exercised. Perhaps the most apparent violation of sound control procedures occurs where one experimenter collects data for a while, after which he is relieved by another experi-

menter, with no plan for assigning an equal number of subjects in each group to each experimenter. But far more subtle effects are also possible. Rosenthal (1966), for instance, has shown that wishes and expectations of the experimenters can actually influence the nature of the data they collect. In one study, for example, an expectation as to how the data should turn out was implanted in 12 experimenters. The experimenters were then assigned to one of three groups. One group of experimenters then obtained data that were in line with their expectations from their first two subjects (who were actually "stooges"); a second group obtained data from their first two subjects (also accomplices) that ran counter to their expectation; the third group of experimenters served as a control condition, and their first two subjects were legitimate and naive. The three groups of experimenters then ran four more individuals, all of whom were legitimate subjects. An analysis of the data for the last four subjects of each group indicated that the experimenters who had received "good" data from their first two subjects also obtained data in line with their expectations from their last four subjects. On the other hand, those experimenters who obtained "bad" data from their first subjects received "bad" data from their last four. Such research thus experimentally demonstrates that knowledge of "early returns" can significantly influence the nature of later collected data.

In another experiment Rosenthal informed one group of experimenters that their rats had been bred so that they were "maze-bright" ("intelligent"). A second group of experimenters were told that they had "maze-dull" rats. The expectation implanted in the experimenters, then, was that the former would be able to quite readily learn a simple discrimination problem, while the learning of the latter animals would be slow. The results showed that experimenters who believed that their subjects were bright obtained performance from them that was significantly superior to that obtained by experimenters who believed that their subjects were dull, even though both groups of rats were of equal intelligence.

It is indeed promising that we are at last systematically investigating the experimenter as a stimulus object in the experimental situation. Even though much remains to be learned, we can at least attempt to take precautions where such are feasible. Clearly, balancing of subjects across experimenters (p. 132) is called for when more than one data collector is involved. The technique of elimination could possibly be resorted to, in which case subjects could be run entirely by means of automated equipment such as tape recorders. In some cases it might be possible to keep knowledge of the nature of the hypothesis, or of the data actually collected, away from the experimenter. A rather extreme technique is cited by Rosenthal (1964): one experimenter (Rosenhan) conducted an experiment, then had it repeated by another experimenter in whom the opposite hypothesis had been ingeniously implanted. Or, one could add additional control groups which

would be labeled "... 'expectancy control groups.' In any study employing an experimental (treatment) and a control (no treatment) condition, a group would be added for whom the experimenter(s) is reasonably led to expect the same sort of data as is expected from the treatment group but in which the treatment is not administered. Differences between the treatment group and the expectancy control group might then be attributable to the treatment truly or to a treatment and expectancy interaction rather than to expectancy alone" (Rosenthal, 1964, p. 111).

Even though such precautions as these might be difficult to take in your elementary experimentation, it is at least valuable for you to become aware of the problem. If you later attempt to conduct publishable research, you will then be sufficiently sophisticated to effect the best controls that you can.

SOME CONTROL PROBLEMS

In the following experiments you should attempt to determine what the control problems are and to specify how you would apply the appropriate techniques to solve the problems. After considering the various experiments you should then reach a conclusion as to whether or not they should have been conducted. Should you like additional practice on this important topic, study the problems offered by Underwood (1966a).

To set the tone for this section we would like to discuss an experiment in which the control, if such existed, was outlandish. One day, a general called the author to say that he was repeating an experiment that the author had conducted on rifle marksmanship, and asked if the author could visit him for the purpose of discussing the experiment. The trip was made and the general immediately drove out to the rifle range where the experiment was in progress. We visited the experimental group to observe their progress. During the visit it was more enjoyable watching the general than the subjects, who were newly "enrolled" army trainees. For while the trainees were practicing firing, the general would walk along the line, kicking some into the proper position, lying down beside others to help them fire, etc. After awhile the general suggested that we leave. That was fine, except for one thing, namely a desire to observe the control group. (By this time the author was beginning to wonder if there was a control group, but this concern was unfounded). The general suggested a walk over the next hill, for that was where the control group was located. On his way the author stopped to talk privately with the sergeant, particularly commenting on how lively and enthusiastic the experimental subjects were. The sergeant explained that that was what the general wanted — that the general expected the experimental group to fire better than the control group and they "darn" well knew that that was what had better happen. When the other side of the hill was reached the author was amazed at the contrast. The control subjects were the most morose,

depressed, laconic group of subjects he had ever seen. In talking to the sergeant in charge of this group he was informed that the general had never been to visit them. What is more, his group knew that they should not perform as well as the experimental group, for nobody wanted the general to be disappointed (their motivations are too numerous to cite here). Needless to say, when the general reported the results of the experiment they were highly significant in favor of the experimental group. Let us now see if you can spot the errors, if any, in the following experiments.

1. The problem of whether children should be taught to read by the Word method or by the Phonics method has been a point of controversy for many years. Briefly, the Word method teaches the child to perceive the word as a whole unit, whereas the Phonics method requires that he break the word into parts. To attempt to decide this issue an experimenter plans to teach reading to two groups, one by each method. The local school system teaches only the Word method. "This is fine for one group," he says. "Now I must find a school system that uses the Phonics method." Accordingly he visits another town that uses the Phonics method. He then decides that he will test a sample of third-grade children in each town to see how well they can read. After administering a long battery of reading tests he finds that the children who used the Phonics method are significantly superior to the children who used the Word method. He then concludes that the Phonics method is superior to the Word method.

2. A military psychologist is interested in whether training to fire a machine gun from a tank facilitates accuracy in firing the main tank gun. He obtains a company of soldiers with no previous firing experience, and randomly divides them into two groups. One group receives .30 caliber machine gun training, the other does not. He then tests both groups on their ability to fire the larger tank gun. To do this he has two tanks set up so that they can fire on targets in a field. The machine-gun-trained group is assigned one tank and a corresponding set of targets, while the control group fires on another set of targets from the second tank. His tests show that the group previously trained on the machine gun is significantly more accurate than the control group. His conclusion is that .30 caliber machine gun training facilitates accuracy on the main tank gun.

3. A psychologist seeks to test the hypothesis that early toilet training leads to a type of personality where children are excessively compulsive about cleanliness; conversely, late toilet training leads to sloppiness. The psychologist notes that previous studies have shown that middle-class children receive their toilet training earlier than do lower-class children. Accordingly he forms two groups, one of middle-class children and another of lower-class children. He then provides both groups with a finger painting task and records a number of data about their procedures, e.g., the extent to which they smear their hands and arms with paints, whether or not they clean up after

the session, and how many times they wash the paints from their hands. Comparisons of the two groups on these criteria indicate that the middle-class children are significantly more concerned about cleanliness than are those of the lower-class. It is thus concluded that early toilet training leads to compulsive cleanliness whereas later toilet training results in less concern about personal cleanliness.

4. A physiological psychologist seeks to determine a function of the internal part of the brain known as the hypothalamus. He obtains a sample of cats and randomly assigns them to two groups. An operation removes the hypothalamus from all the cats in one group. The second group is not operated on. On a certain behavior test it is found that the operated group is significantly deficient, as compared to the control group. The psychologist concludes that the hypothalamus is responsible for the type of behavior that is "missing" in the group that was operated on.

5. The following hypothesis is subjected to test: emotionally loaded words (e.g., "sex," "prostitute") must be exposed for a longer time to be perceived than words that are neutral in tone. To test this hypothesis various words are exposed to subjects for extremely short intervals. In fact, the initial exposure time is so short that no subject can report any of the words. The length of exposure is then gradually increased until each word is correctly reported. The length of exposure necessary for each word to be reported is recorded. It is found that the length of time necessary for subjects to report the emotionally loaded words is longer than for the neutral words. It is concluded that the hypothesis is confirmed.

THE INDEPENDENT AND DEPENDENT VARIABLES

From one point of view, the primary purpose of an experiment is to test a hypothesis. And a hypothesis, we said, is a statement to the effect that two (or more) variables are related. We have referred to the two variables as the independent and the dependent variables. In this chapter we will discuss these variables in greater detail, and also the types of relationships that many obtain between them.

TYPES OF RELATIONSHIPS STUDIED IN PSYCHOLOGY

In general approach, we may develop an analogy between the way an engineer looks at a "machine" such as a computer and the way that a psychologist looks at an organism. For the electronic computer, the engineer first has to put some type of energy into it (he calls this the "input"). The input then activates the computer in such a way that the energy is "carried"

through it (this is the “throughput”). And, finally, the computer accomplishes the task for which it is built and certain actions result (the “output”). There are certain relationships among the input, throughput, and output so that for certain types or amounts of input, certain types and amounts of throughput occur, and certain types and amounts of output result. Furthermore, the characteristics of the computer limit the nature of the throughput and thus of the output. For example, only certain types or amounts of input are capable of being transmitted by the computer. And the characteristics of the computer determine what kinds of output may occur.

The psychologist's approach to behavior is analogous.¹ For he may consider that the organism corresponds to the computer; the stimuli that excite the organism's receptors are the input, and the organism's responses are the output. The analogy may be pursued in the following manner: the type of stimuli that enter the organism (the input) determine what will happen within the organism (the throughput); and what happens within the organism influences the nature of the responses (the output). Furthermore, the specific characteristics of the organism, in particular the neural connections, the past experience, the genetic makeup, also determine the nature of the organism's responses (see Figure 7.1). For example, if a light is flashed in an organism's eye (input), various neural pathways are excited which go to and from the visual areas of the brain (throughput), and thence to specific effec-

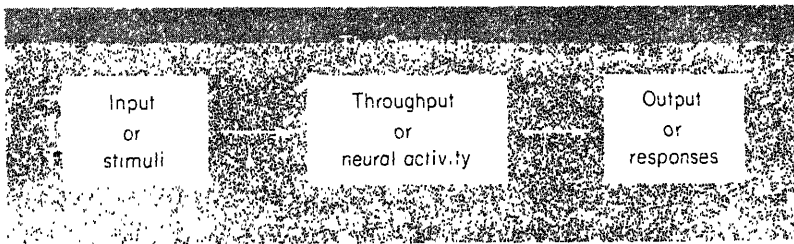


FIGURE 7.1.

An analogy between the approach of an engineer to the study of a computer and the psychologist's approach to behavior.

¹Actually, it would be more interesting to consider the psychologist's terms as a special case of, rather than an analogy to, the input-throughput-output schematization. Or, an even more radical suggestion would be to replace the psychologist's terms with those of the engineer. The implication is an important one for the generality of psychological laws and for the hierarchical status of psychology relative to other sciences. For, assuming the classical definitions of *stimulus* ("energy that excites an organism's receptors") and *response* ("the contraction of muscles or secretion of glands"), it can be noted that the psychologist's terms are relatively limited, particularly as it is often implicitly assumed that they refer to the apparatus of mammals. The use of the more general terms ("input," etc.) might facilitate our search for laws of greater generality, laws that might well apply to lower organisms with unique receptors and effectors, to all manner of "machines" such as electronic computers, and even to extraterrestrial organisms who, in all probability, have very strange receptors and effectors indeed.

tors which result in a given response (output). However, in working with an organism whose visual areas in the brain have been destroyed, different characteristics will be encountered, and a different (or perhaps no) response would occur.

There are, then, three general classes of variables with which the psychologist deals: stimulus variables (the input), organismic variables (the throughput), and response variables (the output).² The psychologist attempts to determine relationships between these three. The possible relationships that may be studied are shown in Figure 7.2.

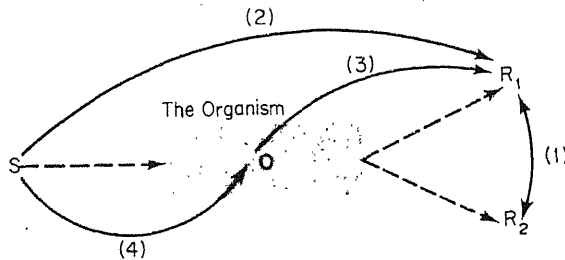


FIGURE 7.2.

Showing the possible relationships among three classes of variables studied in psychology. *S* denotes the stimulus variables; *O*, organismic variables; *R*, response variables. The relationships indicated by numerals 1, 2, 3, and 4 are discussed in the text. (Modified from K. W. Spence, 1948.)

These possible relationships may be indicated symbolically as follows:

1. $R_1 = f(R_2)$ — response number one (any given response) is a function of response number two (any other response).³ When determining that two classes of responses are related, you are determining the first type of relationship. It is, however, difficult to experimentally establish a relationship between two responses, and the application of correlational techniques is a more appropriate approach to this problem. For example, an experimenter may want to know whether two dependent variables are related. In running rats in a maze, for instance, he wants to find out whether the number of errors that rats make is related to the total time it takes the rats to run the maze. These are two response measures, and the correlation between them may be computed.

²Numerous classifications of variables with which psychologists deal are available elsewhere, e.g., Spence (1948) considers stimulus, organic, response, and hypothetical state variables; Underwood (1957) discusses environmental, task, instructional, and subject variables; Edwards (1968) uses the present classification as do Woodworth and Schlosberg (1955) though the latter add antecedent variables to the list.

³Let it be assumed that when we refer to a certain response we mean a certain response class, a number of quite similar instances of responses. For example, a response instance would be one hit at a baseball, whereas a response class would be made up of all of the times that we hit a baseball (a large number of response instances). Similarly, we may refer to stimulus instances and stimulus classes.

2. $R = f(S)$ — a certain response class is a function of a certain stimulus (class). In this case, one may vary values of the stimulus to see if values of the response change. The stimulus may thus be seen to be the independent variable, while the response is the dependent variable. This second type of relationship is that with which we are most often concerned in experimentation. Probably the clearest examples of the areas of psychology in which this type of relationship is sought are those of perception and learning. In studies of perception we vary stimulus conditions and determine whether the perceptual responses of the organism also vary. For instance, we might vary the lighting conditions on a given object (varying the stimulus variable) and see if a person's verbal report of its size changes (a response measure).

3. $R = f(O)$ — a response class is a function of (a class of) organismic variables. The primary purpose of research aimed at this type of relationship is to determine whether certain characteristics of the organism lead to certain types of responses. We might wonder, for instance, if people who are short and stout behave differently than people who are tall and thin. More specifically, do these two types of people differ as far as happiness, general emotionality, or degree of verbosity is concerned?

4. $O = f(S)$ — a class of organismic variables is a function of a class of stimulus variables. In this case, we are primarily asking what environmental events influence the characteristics of organisms. For example, we might be interested in whether the degree to which a child has been isolated influences his intelligence.

This is a brief survey of the basic types of relationships sought by psychologists.⁴ We should add that more complex relationships may also be sought, as for instance those that would occur if you investigate the relationship among three responses [$R_1 = f(R_2, R_3)$], among two stimuli and a given response [$R = f(S_1, S_2)$], or among a stimulus, response, and organismic variable [$R = f(O, S)$]. We would also note that the statement of the types of relationships sought depends on the way that you classify variables. Hence, other systems of classification would lead to different statements of the possible relationships.

THE INDEPENDENT VARIABLE

In the first type of relationship we would vary one response to see if another is thereby affected. The response that we may vary would be our independent variable, the other response our dependent variable. However, as we said in presenting the first type of relationship, response-response relation-

⁴You should note that these relationships conform to our discussion in Chapter Three of the nature of hypotheses. There we showed how a hypothesis stated as a mathematical function [e.g., $R = f(S)$] is a special case of the more general "if a then b " relationship.

ships are not often sought with the use of standard experimental designs. In the second type of relationship we vary a stimulus and determine its effect on a response. Hence, the stimulus is the independent variable, the response the dependent variable. In the third type of relationship we vary an organismic variable as the independent variable and determine its relationship to a response, the dependent variable. And in the fourth type of relationship a stimulus is the independent variable, and an organismic variable is the dependent variable. Thus, we have three independent variables to consider: responses, stimuli, and organismic variables. And in general, the symbol to the right of each equation is the independent variable, that to the left the dependent variable.

RESPONSE VARIABLES

Because of the infrequent use of this type of independent variable in experimentation we shall not discuss it further, except to provide an example. Let us say that we are interested in whether people who read on the subway are steadier than people who do not read on the subway. It may have occurred to some of you who are both people watchers and subway riders that the reason that some people don't read on the subway is that they can't hold newspapers steady enough. Our procedure would be straightforward, assuming we could get the cooperation of our subjects: administer a suitable steadiness test to a group of subway readers and to a group of subway non-readers. The difference, or lack of a difference, between the two groups would answer our question.

STIMULUS VARIABLES

Most independent variables with which we are concerned are stimulus variables, where the word stimulus is used in a broad sense to refer to any aspect of the physical or social environment that excites the receptors. The following are examples of stimulus variables as they might affect a particular kind of behavior: the effect of different sizes of type on reading speed; the effect of different styles of type on reading speed; the effect of intensity of light on the rate of conditioning; the effect of number of people present at dinner on the amount of food eaten; the effect of social atmosphere on problem-solving ability. You will note that the variables differ considerably in their complexity.

ORGANISMIC VARIABLES

By organismic variables we mean any relatively stable characteristic of the organism, including such physical or physiological characteristics as sex,

eye color, height, weight, and body build, as well as such psychological characteristics as intelligence, educational level, anxiety, neuroticism, and prejudice. Perhaps we should also include here the administration of various drugs, for psychopharmacological research is increasing in its importance.

At this point, let us remark that this system of classification, like any other, has its disadvantages, since it is difficult to force some variables into categories. For example, we might question the placement of intelligence in the category of organismic variables. For while it can well be considered to be a characteristic of the organism, let us observe the way in which intelligence is frequently measured: A person takes a pencil and makes a number of marks on a piece of paper — he makes responses. Hence, it is also possible to classify intelligence as a response variable.⁵ We must, therefore, be quite arbitrary in some of our decisions, the justification being that we are trying to consider independent variables in an orderly fashion that will allow some systematic insight into the various kinds of variables used in psychological research.

A FURTHER CONSIDERATION OF INDEPENDENT VARIABLE CONTROL

In Chapter Six we said that independent variable control occurs when the researcher varies the independent variable in a known and specified manner. There are essentially two ways in which an investigator may exercise independent variable control: (1) purposive manipulation of the variable; and (2) selection of the desired values of the variable from a number of values that already exist. When purposive manipulation is used, we say that an experiment is being conducted; but when selection is used, we say that the method of systematic observation is being used. If an experimenter is interested in whether the intensity of a stimulus affects the rate of conditioning, he might vary intensity in two ways, high and low. If the stimulus is a light, he might choose the values, say, two and twenty candle power. He would then, at random: (1) assign his sample of subjects to two groups; and (2) randomly determine which group would receive the low intensity stimulus, which the high. In this case, he is *purposely manipulating* the independent variable (this is an experiment), for the decision as to what values of the independent variable to study and, more important, which group receives which value is *entirely up to him*. And, what is perhaps equally important, the experimenter himself "creates" the values of the independent variable.

Now, let us illustrate *independent variable control by selection of values* as they

⁵To further complicate this decision, it would also be possible to consider intelligence as a logical construct, or in Spence's system of classification, as a hypothetical state variable (see note, p. 146).

already exist (this is the method of systematic operation). Suppose that an investigator is interested in the effect of intelligence on problem-solving. Further, assume that he is not interested in studying the effects of minor differences of intelligence, but wants to study widely differing values of this variable. For example, he might wish to study three values: an IQ of 135, a second of 100, and a third of 65. Up to this point, the procedures in exercising the two types of independent variable control are the same; the investigator determines what values of the variables he wants to study.

However, in this case, the investigator must find certain groups that yield the values of intelligence that he wants. To do this he might administer a number of intelligence tests at three different places. First, he might select a number of bright college students, providing a group that has an average IQ of 135. Second, he might visit a rather nonselective group (high school students or army personnel) to obtain an average value of 100. And third, he might find a mental institution that would yield a group with an average IQ of 65. With these three groups constructed, he would then administer his test of problem-solving ability and reach the appropriate conclusion. Observe that he has selected the values of his independent variable from a large population. *The IQs of the people tested determined who would be the subjects. He has not, as in the preceding example, determined which subjects would receive which value of the independent variable.* Rather, in selection it is the other way around; the value of the independent variable determines which subjects will be used. It is thus apparent that in independent variable control by selection of values as they already exist in subjects, *the subjects are not randomly assigned to groups.*

It is not really practical, however, to settle on precise IQ values, as above. What the experimenter is more likely to do is to say that "I want a very high intelligence group, a medium group, and a very low intelligence group." He might then make his three visits and settle for whatever IQs he gets — in this case, the averages might be 127, 109, and 72, which would probably still accomplish his purpose.

In short, *purposive manipulation* occurs when the investigator determines the values of the independent variable, "creates" those values himself, and determines which group of subjects will receive which value. *Selection* occurs when the investigator selects subjects who already possess the desired values of the independent variable.

The distinction between both kinds of independent variable control is important. To understand this, focus on the intelligence-problem-solving example. What would be the investigator's appropriate conclusion? (Before hypothesizing possibilities, be sure to recall the chapter on control.) Consider the *confounded* nature of this investigation. We have three groups of subjects whom we know differ in intelligence. But in what other respects might they differ? The possibilities are so numerous that we shall only list three: socioeconomic status, the degree of stimulation of their environments, and motiva-

tion to solve problems. Hence, whatever the results on the problem-solving tests, the confounding of our independent variable with extraneous variables would be atrocious. We would not know to which variable, or combination of variables, to attribute possible differences in dependent variable scores. This is not so with an experiment like the light-conditioning example above. For in that case whatever the extraneous variables might be, they would be randomized out — distributed equally — over all groups.

When a stimulus variable is the independent variable, *purposive manipulation* is generally used. If the independent variable is either a response or organismic variable, however, *selection* is the more likely independent-variable-control procedure. For example, with intelligence (or number of years of schooling, or chronic anxiety, etc.) as the independent variable, we have no practical alternative but to select subjects with the desired values. It might be possible to manipulate purposively some of these variables, but the idea is impractical. It is admittedly difficult, say, to raise a person in such a way (manipulating his environment, administering various drugs, etc.) that he will have an IQ of the desired value; we doubt that you would try it.

A number of studies have been conducted to determine whether or not there is a relationship between cigarette smoking and lung cancer. Generally, two groups are studied, one composed of people who do not smoke, the second of those who do. The independent variable is thus the degree of smoking. Measures on the dependent variable are then taken — frequency of occurrence of lung cancer. The results have been generally decisive in that smokers more frequently acquire lung cancer than do nonsmokers. Nobody can argue with this statement. However, the additional statement is frequently made: *Therefore*, we may conclude that smoking causes lung cancer. On the basis of the type of evidence presented above, such a statement is unfounded, for the type of control of the independent variable that has been used is that of selection of values. Numerous additional variables may be confounded with the independent variable.

The only behavioral approach to determine the cause-effect relationship is to exercise control through purposive manipulation. That is, to select at random a group of subjects who have never been exposed to the smoking habit (e.g., children or isolated cultural groups), randomly divide them into two groups, and randomly determine which group will be smokers, which the abstainers. Of course, the experimenter must make sure that they adhere to these instructions over a long period. As members of the two groups acquire lung cancer, the accumulation of this evidence would decide the question. Unfortunately this experiment will probably never be conducted.⁶ But the

⁶Two points might be added here: (1) The research cited does not say that there is no causal relationship, so the wise person, considering the mathematical expectancy, probably would not want to bet that smoking does *not* cause cancer; and (2) at times one can move from the findings gained from the R-R approach with humans to an S-R approach with animals. Hence, one of the values of animal experimentation is that S-R and O-R experiments can be conducted that are not possible with humans.

main point of this discussion should now be quite apparent: Confounding is very likely to occur when *selection* of independent variable values is used (the method of systematic observation) but is considerably less likely when purposive manipulation is resorted to (experimentation).

In studies involving more than one independent variable, the values of one variable might be purposively manipulated and the values of the other selected as they naturally occur. Such an investigation may be referred to as a *quasi-experiment*.

THE DEPENDENT VARIABLE

MEASURES OF THE DEPENDENT VARIABLE

Generally, we have viewed response measures as the dependent variable in psychological experimentation. The class of response measures is extremely broad (comparable to "stimulus," as discussed previously). By "response measures" we mean to include such diverse phenomena as number of drops of saliva a dog secretes, number of errors a rat makes in a maze, time it takes a person to solve a problem, amplitude of *electromyograms* (electrical signals given off by muscles when they contract), number of words spoken in a given period of time, accuracy of throwing a baseball, and judgments of people about certain traits. But whatever the response, it is best to measure it as precisely as possible. In some experiments great precision can be achieved, and in others the characteristics of the events dictate cruder measures. To enhance our understanding, we shall briefly list some standard ways of measuring responses.

1. *Accuracy of the response.* Several ways of measuring accuracy are possible. For example, we might have a metrical system, such as when we fire a rifle at a target. Thus, a hit in the bullseye might be scored a five, in the next outer circle a three, and in the next circle a one. Another type of response measure of accuracy is to count the number of errors the subject makes. For example, the number of erroneous movements a person makes in putting a puzzle together, or the number of blind alleys a rat enters in running a maze.

2. *Latency of the response.* This is a measure of the time that it takes the organism to begin the response, as in the case of reaction time studies. The experimenter may provide a signal to which the subject must respond. He then measures the time interval between the onset of the stimulus and the onset of the response. Or, in the case of a rat running a maze, the latency might be the time interval between the raising of the start box door and the time the rat's hind feet leave the box.

3. *Speed of the response.* This is a measure of how long it takes the organism to complete its response, once it has started. If the response is a simple one like pressing one of two telegraph keys, the time measure would be quite short. But if it is a complex response, such as solving a difficult problem or assembling a complicated device, the time measure would be long. A measure of the speed of the response in the case of a rat running a maze would be the length of time between his leaving the start box, until he reaches the goal box. To emphasize the distinction between latency and speed measures — latency is the time between the onset of the stimulus and the onset of the response, and speed is the time between the onset and termination of the response.

4. *Frequency of the response.* One might also measure the number of times a response occurs, as in the case of how many responses an organism makes before extinction sets in. If the frequency of responding is counted for a given period of time, the rate of responding can be computed. If a response is made ten times in one minute, the rate of responding is ten responses per minute. The rate gives an indication of the probability of the response — the higher the rate, the greater the probability that it will occur in the situation at some future time. Rate of responding is most often used in experiments involving operant conditioning, e.g., the organism is placed in a Skinner Box and each depression of a lever is automatically recorded on a moving strip of paper.

Additional measures of the response might be level of ability that a subject can manifest (e.g., how many problems of increasing difficulty a subject solves with an *unlimited* amount of time), or the intensity of a response (e.g., the amplitude of the galvanic skin response in a conditioning study). Frequently it is impossible to obtain an adequate measure of the dependent variable with any of these techniques. In this event, it might be possible to devise a rating scale. For example, a rating scale for anxiety might have five gradations: 5 meaning “extremely anxious,” 4 “moderately anxious,” and so on. Competent judges would then mark the appropriate position on the scale for each subject. Or the subjects could even rate themselves.

Objective tests frequently serve as dependent variable measures. For example, you might wish to know whether or not psychotherapy decreases a person's neurotic tendencies, in which case, you might administer a standard test for this purpose. If a suitable standard test is not available, it might be that you could construct your own, such as one student did in developing the “Dollenmayer Happiness Scale.”

These are some of the more commonly used measures of dependent variables. By combining some of the above ideas with your own ingenuity, you should be able to arrive at an appropriate measure of a dependent variable for the independent variable that you wish to study.

SELECTING A DEPENDENT VARIABLE

The experimenter seeks to determine if his independent variable affects a dependent variable. But how does the experimenter determine what dependent variable to measure and record? Behavior is exceedingly complex, and at any given time an organism makes a fantastically large number of responses. Take Pavlov's simple conditioning experiment with dogs. His dependent variable (as it is most frequently cited) was amount of salivation. However, that is not the dog's only response when a conditioned and an unconditioned stimulus are presented. For in addition to salivating, he also breathes at a certain rate, wags his tail, moves his legs, pricks up his ears, and so on. Now, out of this mass of behavior, Pavlov had to select a particular response to measure and record. He might have studied some response other than salivation, a response that might or might not have also been related to his independent variable. Why did he choose salivation?⁷

Presumably every stimulus-independent variable leads to certain responses. The problem of selecting a dependent variable, then, would seem simply to find all the responses that are influenced by a given stimulus-independent variable. But the problem is not quite that simple, and even this answer is not a simple one to follow in practice. Look at the matter in terms of our previous distinction between exploratory and confirmatory experiments. In the exploratory experiment, the experimenter asks himself: "I wonder what would happen if I did this?" He selects some responses to measure to see if they are affected by the independent variable. It is impossible in any practical sense to study all of them; you simply pick and hope. An interesting example of this procedure is offered in the following quotation:

"... The discovery that serotonin is present in the brain was perhaps the most curious turn of the wheel of fate. . . . Several years ago Albert Hofman, a Swiss chemist, had an alarming experience. He had synthesized a new substance, and one afternoon he snuffed some of it up his nose. Thereupon he was assailed with increasingly bizarre feelings and finally with hallucinations. It was six hours before these sensations disappeared. As a reward for thus putting his nose into things, Hofman is credited with the discovery of lysergic acid diethylamide (LSD), which has proved a boon to psychiatrists because with it they can induce schizophrenic-like states at will. . ." (Page, 1957, p. 55).

In this "pick and hope" procedure you can reach two possible conclusions from your data: (1) the independent variable did not affect the particular variable; or (2) it did. In the confirmatory experiment, on the other hand, you have a precise hypothesis that indicates the dependent variable in which you are interested; it specifies that a certain independent variable will influence a certain dependent variable. Your procedure is straightforward,

⁷See Skinner (1953, pp. 52-54, for a brief, interesting answer.

at least in principle. You merely select a measure of the dependent variable in question and see if your hypothesis is probably true or false.

Since the dependent variable has already been specified by the hypothesis, you must be careful to obtain a proper measure of it. That is, you must be sure that the data you record are actually measures of the dependent variable *in which you are interested*. Suppose, for instance, that instead of measuring amount of salivation Pavlov measured the change in color of his dog's hair; the whole concept of conditioning would then have been delayed until the appearance of a shrewder investigator. This is a grotesque example, but more subtle errors of the same type are frequently made in psychological research. One experimenter wishing to study the effect of a certain independent variable on emotionality might select several judges to rate the apparent emotionality of his subjects after the independent variable has been introduced. Whatever his results, he should ask the question: Did the judges actually rate subjects on the basis of emotionality, or did they unknowingly rate them on some other characteristic? It may have been that the subjects were actually rated on "general disposition to life," "intelligence," "personal attractiveness," or whatever. If this actually happened, we could not say that emotionality was really the dependent variable that was measured. This brings us to the first requirement that a dependent variable in a confirmatory experiment must meet; it must be *valid*. The data recorded must actually be measures of the characteristics that the experimenter seeks to measure.

"Now," you might say, recalling our discussion on operational definitions, "if the experimenter defined emotionality as what the judges reported, then that is by definition emotionality — you can't quarrel with that." And so we can't, at least on the grounds that you offered. We recognize that anyone can define anything any way that he wants. You can, if you wish, define the typical four-legged object with a flat surface on top from which we eat "a chair" if you like. Nobody would say that you can't. However, at the same time, we must ask you to look at a social criterion: Is that the name usually used by other people? Obviously the object is usually referred to as a "table." And if you insist on referring to what the rest of us call a "table" as a "chair," nobody should call you wrong, for this is your privilege. However, you will be at a distinct disadvantage when you try to communicate with other people. When you invite your dinner guests to be seated on a table and to eat their food from a chair some very quizzical responses will be evoked.

So the lesson is this: Although you may define your dependent variable as you wish, it is wise to define your dependent variable as it is customarily used — at least, if it is customarily used. And if you are lucky enough to investigate a problem that has a certain widely accepted definition for the dependent variable involved, you should either use that dependent variable or one that correlates highly with it.

Consider dependent variable validity by some additional examples. Suppose you are interested in determining the influence of a given independent variable on problem-solving. You might define your dependent variable as the number of problems of a certain nature solved within a given period of time. At first glance, this *seems* a fine example of a valid dependent variable. And, if the test has a large number of problems and if these problems are arranged in ascending order of difficulty, then it probably *is* valid. But if the test is lengthy and the problems are all easy, you probably would not be measuring problem-solving ability but, rather, reading speed. That is, regardless of the fact that "problems" are contained in the test, those who read fast would get a high score and those who read slowly would get a low score. Clearly this would not be a valid measurement of problem-solving ability (unless, of course, problem-solving ability and reading speed are significantly correlated). Or, to make the matter even simpler, if you construct a very short test composed of extremely easy problems, all the subjects will get a perfect score, unless you are working with feeble-minded individuals or the like. This test is not a valid measure of the dependent variable.

In the example where we were interested in whether rats could learn to run to a white or black goal box (p. 132), the training procedure consisted of feeding them in a certain goal box (say a white one). The test consists of running the rats for a number of trials in a two-choice maze that contains one black and one white goal box. Let us say, for the purposes of making our point, that the white box is always on the right and the black box is always on the left. Assume that the preponderance of runs we record are to the white box. We conclude that the rats run to the box of the color in which they were previously fed. Now, are we really measuring the extent to which they run to the *white* box? Rats have position habits; they frequently are either "left turners" or "right turners" (or they may alternate in certain patterns). If we have selected a group of rats that are all right turners, our measure may be simply of position habits, rather than of the dependent variable we are interested in. Hence, in this example, we are measuring "frequency of turning right" rather than "frequency of running to white."

To determine the validity of the dependent variable, the experimenter might correlate his dependent variable scores with scores obtained by the same subject on some other measure that is known to be valid. If the correlation is high, his measure is valid; if low or not significant, it is not valid. For example, suppose we have a valid measure of anxiety available, but that, for various reasons, we wish to use a different measure of anxiety as our dependent variable. To determine the validity of the measure that we wish to use, we could take both measures on a number of subjects. By computing the correlation between the two sets of scores, we could reach a conclusion as to the validity of our measure. Unfortunately, such a procedure is seldom practical in an experimental situation, primarily because we typically do not

have a measure of the desired dependent variable that has known validity.

Experts in the field of testing have re-examined the classical concept of validity and now recognize several different kinds of validity for their tests. Although problems of validation are still vexing, these specialists have made considerable progress, and some of their concepts and procedures (such as that of construct validity) may well be beneficially applied in experimental psychology. A thorough consideration of the details of validation should be undertaken in a course on testing. We have here merely attempted to bring the topic of validation of the dependent variable to your attention, to point out some of the types of problems present, and to illustrate some of the kinds of pitfalls that await unwary experimenters. As a minimum, you are now aware of the existence of these problems and potential errors. On the basis of considerable reflection and on the basis of results of previous research in any given area, your chances of selecting a valid dependent variable measure should be increased.

The second requirement that a dependent variable should satisfy is that of *reliability*, in part the extent to which subjects receive about the same scores when repeated measurements are taken. For example, an intelligence test may be considered sufficiently reliable if subjects make approximately the same scores every time they take the test. Suppose a subject received an IQ of 105 the first time he takes a test, 109 the second, and 102 the third. These scores are approximately the same. If most subjects behaved similarly, it may be said that the test is reliable for the population sampled. However, suppose that a typical subject scored 109, 138, and 82. Such a test could not be considered reliable, for the repeated measurements vary too much.

Considering an experimental situation, reliability of a dependent variable could be measured in the following manner. First, the experimenter could obtain measures on the dependent variable, preferably from individuals not involved in his experiment. After a period of time, he could test the same subjects on the same measure again. He would then compute the correlation between the two sets of measures. If the correlation is high, his dependent variable measure is reliable; otherwise it is not. Another approach would be to compute a split-half reliability coefficient. Suppose that the dependent variable is a measure that could be divided into two halves. For instance; each subject might be required to solve 20 problems, and the experimenter could therefore obtain a total score for the odd numbered and for the even numbered problems. It would then be a simple matter of computing a correlation coefficient between these two resulting scores for all of his subjects, a value that hopefully would be quite high.

We said above that reliability is in part the extent to which subjects receive the same scores when repeated measurements are taken. To elaborate on this, let us note that experimenters frequently study subject's characteristics that change with the passage of time. They may study a learning process or

the growth of situational anxiety. In such a case we need merely note that the correlation of successive scores may be quite high providing that the subjects maintain the same relative scores: that is, providing that the rank order of scores is similar on each testing. For example, if three subjects made scores of 10, 9, and 6 on the first testing, but 15, 12, and 10 respectively on the second testing, the correlation (reliability) would be high since they maintained the same relative ranks. If, however, the first subject's score changed from 10 to 11, the second from 9 to 12, and the third from 6 to 12, the reliability would be lower. In short, then, whether or not the measures of the dependent variable change with time, a correlation coefficient can be computed to determine the extent to which the dependent variable is reliable.

Unfortunately, most experimenters do not bother to even consider the reliability of their dependent variables. Our knowledge about reliability in experimentation would be increased if they did. At the same time we must point out that the determination of reliability is sometimes unrealistic. There are situations in which the dependent variable is more reliable than a computed correlation coefficient would indicate. For one thing, the subjects used are frequently too homogeneous to allow the computed correlation value to approach the true value. For instance, if all subjects in a learning study had precisely the same ability to learn the task presented them, then (ideally) they would all receive exactly the same dependent variable scores on successive testings; the computed correlation would not be indicative of the true reliability of the dependent variable. Another reason for a different computed correlation than the true one is that the scale used to measure the dependent variable may have insufficient range. To illustrate again by taking an extreme case, suppose that only one value of the dependent variable was possible. In this event all subjects would receive that score, say five, and the computed correlation would again be untrue. More realistically an experimenter might use a five point scale as a measure of his dependent variable, but the only difference between this and our absurd example is one of degree. The five point scale might still be too restrictive in that it does not allow one to sufficiently differentiate among the true scores of his subjects; two subjects for instance might have true scores of 3.6 and 3.9, but on a scale of five values they would both receive scores of 4.0.

Recognizing that it is desirable for the experimenter to determine the reliability of his dependent variable and that it is frequently not feasible to approximate the true value by correlating successive scores, we may ask what the experimenter does. The answer is that he plans the experiment and collects his data. If he finds that his groups differ significantly, he may look back and reach some tentative conclusions about the reliability of the dependent variable. For if his groups differ significantly, this means that they differ to a greater extent than can be accounted for by experimental error.

And if the means on his dependent variable have differed more than can be expected by random fluctuations, it must possess sufficient reliability, for lack of reliability makes for only random variation in scores. On the other hand, if his groups do not differ significantly, this means that the scores are probably due to random variation, to experimental error. The conclusion that would most frequently be reached in such a situation is that variation of the independent variable does not affect the dependent variable. But other reasons are also possible. It may be that the dependent variable is unreliable. So this approach to determining reliability is a one-way affair: If there are significant differences among groups, the dependent variable is probably reliable; if there are no significant differences, then no conclusion about reliability is possible (at least on this information only). The repetition of the experiment a number of times with consistently significant results would increase our belief in the reliability of the dependent variable, for if the same results are continually obtained, certainly the dependent variable is reliable.

The concepts of validity and reliability have been extensively used by test constructors. They have been almost totally ignored by experimenters, and yet their great importance to experimentation should be apparent. If the experimenter has not selected a valid and reliable criterion (dependent variable), his experiment is worthless. If he is performing a learning experiment and his dependent variable actually measures drive, then obviously his conclusions with regard to learning are baseless. The close tie between validity of the dependent variable and experimental control may also be emphasized. If an uncontrolled extraneous variable is affecting the dependent variable, the measures obtained may be valid for the extraneous variable, but not for the independent variable. On the other hand, if the learning experiment has an unreliable dependent variable, then the scores of the subjects, regardless of the experimental conditions, would vary at random. With all of the subject's scores varying in a chaotic manner, it is impossible to determine the effectiveness of the independent variable.

MULTIPLE DEPENDENT VARIABLES

A given independent variable may affect a number of measures of behavior, and in many experiments a number of measures actually are recorded. For example, an experiment in which rats are run through a maze might use all of the following measures of behavior: time that it takes to leave the starting box (latency), time that it takes to enter the goal box (running time or speed), number of errors made on each trial, and number of trials required to learn to run the maze with no errors. Such an experiment could be looked upon as one with four dependent variable measures, in which case the experimenter would merely conduct four statistical analyses; he might run, say, four

separate *t*-tests to see if his groups differ on any of the four dependent variables.

Should this procedure be used, it would be valuable to obtain the correlations among the several dependent variables. You might find that two dependent variable measures correlate quite highly, e.g., .95.⁸ In this case, you would know that they are measuring largely the same thing, and there would be little point in recording both measures in future experimentation. Hence, in later work you would select one or the other, probably the easiest to record. We would emphasize, however, that the correlations between your dependent variables should be quite high before you eliminate any of them. The author once conducted an experiment in which three dependent variables were used. It was found that the correlation between the first and the second was .70, between the first and the third .60, and between the second and the third .80. Yet, in spite of these significant correlations, the statistical analysis indicated that there was no difference between experimental conditions on two of the dependent variables, but that there was a difference significant beyond the one per cent level for the conditions on the third criterion.

In short, then, it is desirable to measure every dependent variable that might reasonably be affected by your independent variable. This statement is offered as an ideal that can only be approached, for in any actual experiment, it would not be feasible. If you have definite information that indicates a high correlation among some of your dependent variables, you might choose among them.

GROWTH MEASURES

More often than not in psychology experimenters deal with variables that change with time. This is universally true with learning studies. For example, we may be interested in how a skill grows with repeated practice under two different methods. Frequently a statistical test is run on terminal data, i.e., data obtained on only the last trial. However, the learning curves of the two groups could provide considerable information about how the two methods led to their terminal points; subjects using one method might have been "slower starters" but gained more rapidly at the end. And, in addition to providing such valuable information, you may want to run statistical tests that compare the two curves at specific points or even as whole units.⁹

⁸See p. 166 for a discussion of correlation and an interpretation of this number. Incidentally, such a correlation should be computed separately for each group in the experiment, i.e., one should not combine all subjects from all groups and compute a general correlation coefficient. For two groups, one would compute two correlation coefficients. In your future study be alerted to the difference between intra and inter class correlations.

⁹This is known as trend analysis, see for example, Edwards (1968) or Lindquist (1953).

DELAYED MEASURES

Another important question concerns the possible retention of experimental effects. For example, we might find that one method leads to better learning than a second method, but we might also be interested in whether that advantage is maintained over a period of time (with various kinds of intervening activity). Suppose that your task is to train mechanics in the performance of a highly technical job. The training you give them is to be followed by training on something else, so that it will be quite a while before they will actually use the training they received from you. In this case, you would not only be interested in which of several methods is more efficient for learning but also which method leads to the best retention. If you conducted an experiment, you might have the men return to you for another test just before they started their on-job duty. On the basis of this delayed test, you could decide which of the several methods would be best to use. Unfortunately, psychologists seldom take delayed measurements of their experimental effects, even when such a practice would be quite easy for them.

EXPERIMENTAL DESIGN

The Case of Two-Matched-Groups

The type of design that we have considered up to this point is the two-groups design, which requires that subjects be assigned to each group in a random fashion. The two-randomized-groups design is based on the assumption that the chance assignment will result in two essentially equal groups (as determined, of course, by comparing their means). The extent to which this assumption is justified, we have also said, increases with the number of subjects used.

The basic logic of all experimental designs is the same: start with groups that are essentially equal, administer the experimental treatment to one and not the other, and note the changes on the dependent variable. If the two groups had equivalent means on the dependent variable before the administration of the experimental treatment, and if a significant difference between the means of the groups on the dependent variable results after the administration of the experimental treatment, and if extraneous variables have been adequately controlled, then that difference on the dependent variable may be attributed to the experimental treatment. The matched-groups design is simply one way of helping to satisfy the assumption that the groups have

essentially equal dependent variable values prior to the administration of the experimental treatment.

A SIMPLIFIED EXAMPLE OF A TWO-MATCHED-GROUP DESIGN

Let us say that we are interested in testing the hypothesis that both reading and reciting material leads to better retention than reading alone. We might form two groups of subjects, one to learn the material by reading and reciting, the second group to spend all their time in reading. If we were using a randomized-groups design, we would assign subjects to the two groups at random, regardless of what we might know about them. With the matched-groups design, however, we use scores on an initial measure called the matching variable to help assure equivalence of groups. Let us say that we use intelligence test scores as our matching variable. We might have ten subjects available; fictitious scores for them are presented in Table 8.1.

Table 8.1. *Fictitious Scores of a Sample of Subjects on a Matching Variable.*

<i>Subject No.</i>	<i>Intelligence Test Score</i>
1	120
2	120
3	110
4	110
5	100
6	100
7	100
8	100
9	90
10	90

Our strategy is to construct the groups so that they are equal in intelligence. To accomplish this we need to pair the subjects who have equal scores, assigning one member of each pair to each group. It is apparent that the following subjects can be paired: 1 and 2, 3 and 4, 5 and 6, 7 and 8, and 9 and 10. The method we shall use for dividing these pairmates into two groups is randomization. This assignment by randomization is necessary in order to prevent possible experimenter biases from interfering with the matching. For example the experimenter may, even though he is unaware of such actions, assign more highly motivated subjects to one group even though each pair has the same intelligence score. By a flip of a coin we might determine that Subject 1 goes in the Reading and Reciting group, Number 2 goes in the Reading group. The next coin flip might determine that Subject 3

goes into the Reading group and Number 4 in the Reading and Reciting group. And so on for the remaining pairs (see Table 8.2).

Table 8.2. *The Construction of Two Matched Groups on the Basis of Intelligence Scores.*

READING GROUP		READING AND RECITING GROUP	
<i>Subject No.</i>	<i>Intelligence Score</i>	<i>Subject No.</i>	<i>Intelligence Score</i>
2	120	1	120
3	110	4	110
6	100	5	100
7	100	8	100
10	90	9	90
	520		520

We may note that the sums (and therefore the means) of the intelligence scores of the two groups in Table 8.2 are equal. Let us now assume that the two groups are subjected to their respective experimental treatments and that we obtain the retention scores for them indicated in Table 8.3 (the higher the score, the better they retained the learning material). Note that we have placed the pairs in rank order according to their initial level of ability on the matching variable, i.e., the most intelligent pair is placed first, and the least intelligent pair is placed last.

Table 8.3. *Fictitious Dependent Variable Scores for Pairs of Subject Ranked on the Basis of Matching Variable Scores.*

INITIAL LEVEL OF ABILITY	READING GROUP		READING AND RECITING GROUP	
	<i>Subject No.</i>	<i>Retention Score</i>	<i>Subject No.</i>	<i>Retention Score</i>
1	2	8	1	10
2	3	6	4	9
3	6	5	5	6
4	7	2	8	6
5	10	2	9	5

STATISTICAL ANALYSIS OF A TWO-MATCHED-GROUPS DESIGN

The two groups of scores in Table 8.3 suggest that the group that both read and recited their material is superior to the reading-only group, but as before, we must ask, are they significantly superior? To answer this question

we may apply the t -test, although the application will be a bit different for a matched-groups design. The equation is:

$$(8.1) \quad t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{n}}{n(n-1)}}$$

The symbols are the same as those previously used, with the exception of D , which is the difference between the dependent variable scores for each pair of subjects. To find D we subtract the retention score for the first member of a pair from the second. For example, the first pair consists of Subjects 2 and 1. Their scores were 8 and 10, respectively, and $D = 8 - 10 = -2$. Since we will later square the D scores (to obtain $\sum D^2$), it makes no difference which group's score is subtracted from which. We could just as easily have said: $D = 10 - 8 = 2$. The only caution to observe is that we need to be consistent, i.e., we must always subtract the Reading group's score from the Reading-Reciting group's score, or vice versa. Completion of the D calculations is shown in Table 8.4.

Table 8.4. *Computation of the Value of D .*

<i>Initial Level of Ability</i>	<i>Reading Group</i>	<i>Reading and Reciting Group</i>	<i>D</i>
1	8	10	-2
2	6	9	-3
3	5	6	-1
4	2	6	-4
5	2	5	-3

Equation (8.1) instructs us to perform three operations with respect to D : First, to obtain $\sum D$, the sum of the D scores, i.e.,

$$\sum D = (-2) + (-3) + (-1) + (-4) + (-3) = -13$$

Second, to obtain $\sum D^2$, the sum of the squares of D , i.e., to square each value of D and to sum these squares as follows:

$$\begin{aligned} \sum D^2 &= (-2)^2 + (-3)^2 + (-1)^2 + (-4)^2 + (-3)^2 \\ &= 4 + 9 + 1 + 16 + 9 = 39 \end{aligned}$$

Third, to compute $(\sum D)^2$, which is the square of the sum of the D scores, i.e.,

$$(\sum D)^2 = (\sum D) (\sum D)$$

Recall that n is the number of subjects in a group (not the total number of subjects in the experiment). In a design where we match (pair) subjects we may safely assume that the number of subjects in each group is the same. In our example $n = 5$. The numerator is the difference between the (dependent variable) means of the two groups, as with the previous application of the t -test. The means of the two groups are 4.6 and 7.2. Substitution of all these values in Equation (8.1) results in the following:

$$t = \frac{7.2 - 4.6}{\sqrt{\frac{39 - \frac{(-13)^2}{5}}{5(5 - 1)}}} = 5.10$$

The equation for computing the degrees of freedom for the matched t -test is: $df = n - 1$. (Note that this is a different equation for df from that for the two-randomized-groups design). Hence, for our example, $df = 5 - 1 = 4$. Consulting our Table of t (p. 108), as before, with a t of 5.10 and 4 degrees of freedom we find that our t is significant at the 1 per cent level ($P < 0.01$). We may thus reject our null hypothesis (that there is no difference between the population means of the two groups) and conclude that the groups differ significantly. If these were real data we would note that the mean for the Reading-Reciting group is higher than that for the Reading group and we would conclude that the former is significantly superior; the hypothesis would be confirmed.

CORRELATION AND THE TWO-MATCHED-GROUPS DESIGN

THE MEANING OF CORRELATION

To adequately understand this design we shall have to consider the topic of correlation. A correlation is a measure of the extent to which two variables are related. The measure of correlation that we shall be concerned with is symbolized by r . This symbol stands for the Pearson Product Moment Coefficient of Correlation. Since the value of r tells us the extent to which two variables are (linearly) related, it is a very valuable statistic that is used in a variety of ways. The value that r may assume varies between $+1.0$ and -1.0 . A value of $+1.0$ indicates a perfect positive correlation and -1.0 indicates a perfect negative correlation. To illustrate this let us say that a group of people have been administered two different intelligence tests. Since both tests presumably measure the same thing, we may assume that the scores are highly correlated. They might be as indicated in Table 8.5.

We may note that the subject who received the highest score on Test A also received the highest score on Test B. And so on down the list, Subject 6 receiving the lowest score on both tests. A computation of r for this very

Table 8.5. *Fictitious Scores on Two Intelligence Tests Received by Each Subject.*

Subject No.	Score on Intelligence Test A	Score on Intelligence Test B
1	120	130
2	115	125
3	110	120
4	105	115
5	100	110
6	95	105

small sample would yield a value of $+1.0$. Hence, the scores on the two tests are perfectly correlated; notice that whoever is highest on one test is also highest on the other test, whoever is lowest on one is lowest on the other, and so on with *no exception being present*.¹ Now let us say that there are one or two exceptions in the ranking of the test scores. Suppose that Subject 1 had the highest score on Test A but the third highest score on Test B; that Number 3 had the third highest score on Test A but the highest score on Test B; and that all other relative positions remained the same. In this case the correlation would not be perfect (1.0) but would still be rather high (it would actually be .77).

Moving to the other extreme let us see what a perfect negative correlation would be, i.e., one where $r = -1.0$. We might administer two tests, one of democratic characteristics and one that measures amount of prejudice (see Table 8.6). The person who scores highest on the first test receives the

Table 8.6. *Fictitious Scores on Two Personality Measures for Each Subject.*

Subject	Score on Test of Democratic Characteristics	Score on Test of Prejudice
1	50	10
2	45	15
3	40	20
4	35	25
5	30	30
6	25	35

lowest score on the second. This inverse relationship may be observed to hold for all subjects without exception, resulting in a computed r of -1.0 . Again if we had one or two exceptions in the inverse relationship the r would be something like $-.70$ indicating a negative relationship between the two variables, but one short of being perfect.

¹Actually another characteristic of the scores must also be present for this type of correlation to be perfect. That is that the interval between successive pairs of scores on one variable must be proportional to the corresponding pairs on the other variable. In our example five IQ points separate each subject on each test. However, this requirement is not crucial to the present discussion.

To summarize, given measures on two variables for each subject a positive correlation exists if, as the value of one variable increases, the value of the other one also increases. If there is no exception the correlation will be high and possibly even perfect; if there are relatively few exceptions it will be positive but not perfect. Thus, as test scores on intelligence Test A increase the scores on Test B also increase. On the other hand, if the value of one variable decreases while that of the other variable increases, a negative correlation exists. No exception indicates that the negative relation is high and possibly perfect. Hence as the extent to which people exhibit democratic characteristics increases the amount of their prejudice decreases; and this, of course, is what we could expect.

One final point concerning the value that r may assume. If $r = 0$ one may conclude that there is a total lack of (linear) relationship between the two measures. In other words as the value of one variable increases the value of the other variable varies in a random fashion. Examples of situations where we would expect r to be zero would be where we would correlate height of forehead with intelligence, or number of books that a person reads in a year with the length of his toenails.² Additional examples of positive correlations would be the height of a person and his weight, his IQ and his ability to learn, and his grades in college and his high school grades. We would expect to find negative correlations between the amount of heating fuel a family uses and the outside temperature, or the weight of a person and his success as a jockey.

In science we seek to find relationships between variables. And a negative relationship (correlation) *is just as important* as a positive relationship. Do not think that a negative correlation is undesirable or that it indicates a lack of relationship. To illustrate, for a fixed sample, a correlation of $-.50$ indicates just as strong a relationship as a correlation of $+.50$, and a correlation of $-.90$ indicates a stronger relationship than does one of $+.80$.

THE COMPUTATION OF A CORRELATION COEFFICIENT

The most frequently used correlation coefficient is the Pearson Product Moment Coefficient of Correlations, and it is symbolized by r .³ However, we

²It has been argued that this would actually be a positive correlation on the grounds that excessive book reading cuts into a person's toenail cutting time. Resolution of the argument must await relevant data.

³An equation for computing r directly from raw data is:

$$r_{XY} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

Where r_{XY} is the correlation between two variables X and Y and ΣXY is the sum of the cross products of the values of X and Y for each subject.

shall illustrate the computation of a different, but related, correlation coefficient — it is called The Spearman Rank Correlation Coefficient, which is symbolized by r_s . r_s has the advantage that it is quicker and easier to compute, and it can be conveniently computed without a calculator. By and large, what we have said for r is true for r_s — the only difference of note is that the Spearman Rank Correlation Coefficient is slightly less powerful than the Pearson r . The equation for computing a Spearman-Correlation Coefficient is:

$$(8.2) \quad r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

We shall illustrate the computation of r_s by using the fictitious data of Table 8.6. Equation (8.2) tells us that there are two basic values that we need: (1) d , which is the difference between the ranks of the two measures that we are correlating; and (2) n , which is the number of subjects in the sample. The value of n is, obviously, six, so let us turn to d . To compute d , we need to rank order the scores for each variable separately, i.e., we assign a rank of one to the highest score for the first variable, a rank of two to the second highest score for that variable, and so on. Then we similarly rank the scores for the second variable. The ranks for the two variables of Table 8.6 are presented in Table 8.7.

Table 8.7. *Ranks of the Scores in Table 8.6 and the Computation of d^2 .*

Subject	Rank on Test of Democratic Characteristics	Rank on Test of Prejudice	d	d^2
1	1	6	-5	25
2	2	5	-3	9
3	3	4	-1	1
4	4	3	1	1
5	5	2	3	9
6	6	1	5	25
				$\Sigma d^2 = 70$

We can note, for example, that Subject 1 scored the highest on the test of democratic characteristics and the lowest on the test of prejudice. He thus received ranks of one and six on these two tests. To compute d we subtract his second rank from his first, i.e., $1 - 6 = -5$. “-5” is thus entered under the column labeled “ d .” And so on for the other differences in ranks for the remaining subjects. The value of d is then squared in the final column, and the sum of the squares of d is entered at the bottom of the column, viz.,

$\Sigma d^2 = 70$. We are now ready to substitute these values into Equation (8.2) and to compute r_s :

$$\begin{aligned} r_s &= 1 - \frac{6(70)}{6^3 - 6} \\ &= 1 - \frac{420}{216 - 6} \\ &= 1 - 2 = -1.00 \end{aligned}$$

And, as we already knew, these two arrays of scores are perfectly, though negatively, correlated. You are now in a position to quickly compute a Spearman Rank Correlation Coefficient between any two sets of scores that are of interest to you.

SELECTING THE MATCHING VARIABLE

Recall that in matching subjects we have attempted to equate our two groups with respect to their mean scores on the dependent variable. In other words, we have selected some initial measure of ability by which to match the subjects and have assigned them to two groups on the basis of these scores so that the two groups are essentially equal on this measure. If the matching variable is highly correlated with the dependent variable scores, our matching has been successful.⁴ For in this event we largely equate the groups on their dependent variable scores by using the indirect measure of the matching variable. If the scores on the matching variable and the dependent variable do not correlate to a noticeable extent, however, then our matching is not successful. In short, the degree to which the matching variable values and the dependent variable values correlate is an indication of our success in matching.

How can we find a matching variable that correlates highly with our dependent variable? It might be possible to use the dependent variable itself. For example, if we are studying the process of learning to throw darts at a target, there might be two methods of throwing that we wish to evaluate. The design would call for one group of subjects to use Method A, the other Method B. To assign subjects to the groups by matching, we might first have all subjects throw darts for five trials. We could use their scores on these five trials as the basis for pairing them off into two groups. They would then be trained by the two methods, respectively, and a later proficiency

⁴Let us emphasize that r (and r_s) is a measure of the extent to which two variables (in our case the matching and the dependent variables) are linearly related. We are, thus, simplifying our discussion by considering only linear relationships between our two variables. Curvilinear relationships are excluded from the above discussion because our knowledge about the several possible correlations involved in Equation (8.4) (p. 180) is considerably limited.

score computed. The *t*-test for matched groups would then be run on that later proficiency score. Our matching would be judged successful to the extent that the first set of scores correlated with the later set of scores. Since both sets of scores are obtained on the same task, we would expect the correlation to be rather high. Thus, it is clear *that the matching variable that is most likely to show a correlation with the dependent variable is that dependent variable itself.*

However, it should be apparent that this technique is not always feasible. Suppose that the dependent variable is a measure of rapidity in solving a problem. If practice on the problem is first given to obtain matching scores, then everyone would know the answer when it is administered later as a dependent variable. Or take another example where we create an artificial situation to see how people react under stress. Using the same situation to take initial measures for the purpose of matching subjects would destroy the novelty of the situation after the independent variable is administered. In such situations we must find other measures that are highly correlated with dependent variable performance.

In the problem-solving example we might give the subjects a different, but similar problem to solve and match on that. Or, if our dependent variable is a list of problems to solve, we might select half that list to use as a matching variable and use the other half as a dependent variable. In the stress example it might be reasonable to assume that a psycho-physiological measure of stress would be related to performance during stress. For example we might take a measure of how much the subjects sweat under normal conditions and assume that those who normally sweat a lot are highly anxious individuals (cf. Mowrer, 1953, pp. 591-640). Matching on such a test might be feasible.

A widely used matching variable in human learning studies is a measure of intelligence. The assumption is that the higher the intelligence the better the learning capacity. Intelligence test scores are quite easy to obtain or may perhaps already be on file in the case of college students.

Another general possibility is to match subjects on more than one variable. In a learning experiment, we might match subjects on initial learning scores and intelligence. Further consideration might suggest additional measures that could be combined with these two.⁵

Now we have said that if a matching variable does not correlate rather highly with the dependent variable, a matched-groups design should probably not be used. For this reason you should be rather certain that a high correlation exists between both variables. You might consult previous studies, for they may provide information on correlations between your depen-

⁵However it is frequently advisable to use a special technique for combining the various measures, discussion of which is probably not too fruitful here. For further information on one way to use more than one matching variable you are referred to Peters and Van Voorhis (1940).

dent variable and various other variables. You could then make a selection from among those that correlate most highly. Of course, you should be as sure as possible that the same correlation holds for your subjects with the specific techniques that you use.

You might also conduct a pilot study where you would take a number of measures on some subjects including your dependent variable measure. Selection of the measure with the highest correlation with the dependent variable would afford a fairly good criterion, if it is sufficiently high. If it is too low, you should pursue other possibilities, or consider abandoning the matched-groups design.

One procedural disadvantage of matching occurs in many cases. When using initial trials on a learning task as our matching variable, we need to bring the subjects into the laboratory to obtain the data on which to match them. Then, after computations have been made and the matched groups formed, the subjects must be brought back for the administration of the independent variable. The requirement that subjects be present twice in the laboratory is sometimes troublesome. It is more convenient to use measures that are already available, such as intelligence test scores or college board scores. It is also easier to administer group tests, such as intelligence or personality tests, which can be accomplished in the classroom. On the basis of such tests appropriate subjects can be selected and assigned to groups before they enter the laboratory.

A MORE REALISTIC EXAMPLE

The example of a matched-groups design and its statistical analysis that we previously used was constructed so that we could "breeze through" it in order to observe the general principles involved. There are, however, a number of details that prove somewhat troublesome when using this design, so let us illustrate it with a more realistic problem.

The data that we shall use were taken from a study in which a principle from S-R theory was subjected to test (McGuigan, Calvin and Richardson, 1959). The aspect of the experiment with which we shall be concerned dealt with performance of subjects at difficult and easy choice points in a stylus maze. The subjects were blindfolded and required to learn a maze (for the hand) that contained ten choice points. Their stylus, held in the hand, was placed at the starting point, and the subjects progressed through the maze until they arrived at the end. Each time they took the wrong path at a choice point, an error was scored. The subjects continued to practice the maze until they learned it perfectly. The five easiest and the five most difficult choice points had previously been determined, and the number of errors made by each subject for each of these two categories was counted. The

principle from S-R theory that was tested made a prediction about the performance of the subjects as a function of the difficulty of the choice points. More particularly, it said that the subjects with high drive should not perform as well (make more errors) at the difficult choice points as subjects with low drive. The drive level of the subjects was measured by administering the Taylor Manifest Anxiety Scale (MAS), (Taylor, 1953). The subjects with the highest anxiety level were the high-drive subjects, and those with the lowest anxiety levels were the low-drive subjects. In short, two groups of subjects were formed — high- and low-drive subjects. And it was predicted that the high-drive subjects would make more errors at the difficult choice points than would the low-drive subjects. There was, however, one final qualification for selecting subjects for the particular aspect of the experiment with which we shall be concerned — that the two groups of subjects did not differ as far as learning ability was concerned. That is, it was predicted that differences in drive (anxiety) would lead to differences in performance, so to ascertain that any obtained differences in performance were due to variation of drive, learning ability should be held constant. Equalization of learning ability was accomplished by selecting pairs of subjects in the high- and low-drive groups who made the same total number of errors in learning the maze.

With this general understanding of the rationale of the experiment, which was ingeniously developed from general S-R theory by Farber and Spence (1953) let us progress in a more specific manner through each step. First, it was necessary to measure drive level, so 56 subjects were administered the Taylor Manifest Anxiety Scale. These subjects then practiced the maze until they learned to progress through it with no errors, during which time the number of errors made at each choice point was tallied. To select the specific high- and low-drive subjects used, we shall consider the ten subjects who had the highest MAS scores and the ten subjects who had the lowest. Table 8.8 presents the MAS scores and the total number of errors for these two classes of subjects.

Table 8.8. *Anxiety Scores and Total Numbers of Errors to Learn the Maze for High- and Low-Drive Subjects.*

HIGH-DRIVE SUBJECTS			LOW-DRIVE SUBJECTS		
No.	MAS Score	Number of Errors	No.	MAS Score	Number of Errors
1	36	11	11	1	17
2	35	18	12	4	67
3	35	44	13	6	10
4	33	26	14	7	18
5	30	6	15	7	20
6	29	13	16	8	28
7	29	12	17	8	14
8	28	11	18	10	12
9	28	21	19	10	63
10	28	5	20	10	28

Now, having ascertained high- and low-drive groups, we next need to pair the subjects from each group according to the total number of errors that they made. This task well illustrates why the present is a "more realistic" example than the previous one. To proceed, let us consider Subject 1 who made 11 errors. Who in the low-drive group should this subject be paired with? None of the low-drive subjects made precisely this number of errors, but we can note that Subject 13 made 10 errors and that Subject 18 made 12 errors; either of these two subjects would be satisfactory (though not perfect) as a pairmate. Subject 2 can be perfectly matched with Subject 14, for they both made 18 errors. When we look at Subject 3, who made 44 errors, we can find no reasonable pairmate and thus exclude that subject from further consideration. By further examining the subjects in this manner, the original experimenters finally arrived at five pairs of subjects who were satisfactorily matched; there was no "mismatch" of more than one error. The remaining 10 subjects could not be matched in a reasonable manner, and thus were not studied further. The resulting matched groups are presented in Table 8.9.

Table 8.9. *High- and Low-Drive Groups Matched on Total Numbers of Errors. Pairs of subjects are ranked according to number of errors.⁶*

HIGH-DRIVE SUBJECTS			LOW-DRIVE SUBJECTS		
No.	MAS Score	Number of Errors	No.	MAS Score	Number of Errors
9	28	21	15	7	20
2	35	18	14	7	18
6	29	13	17	8	14
7	29	12	18	10	12
1	36	11	13	6	10
$\bar{X} = 15.00$			$\bar{X} = 14.80$		
$s = 4.30$			$s = 4.15$		

By excluding ten subjects we have been able to achieve a good matching between the two groups, as seen by comparing their means.⁷ Incidentally, we may note one difference between this and the previous example as far as matching is concerned. In the previous example we paired subjects and randomly determined which of each pair went in which group. In the present example, however, groups were formed on the basis of a personality characteristic; the MAS score determined to which group they were assigned.

⁶ s stands for standard deviation, and is discussed on p. 181.

⁷But not without some cost. For by discarding subjects we are possibly destroying the representativeness of our sample. Hence the confidence that we can place in our generalization to our population is reduced. We might also add that we would be interested in comparing the groups on the basis of a measure of variability of the scores. In this case the groups would be rather well matched with regard to their variability as evidenced by the standard deviations of 4.30 and 4.15 for the high- and low-drive groups respectively.

Either procedure is legitimate, and we simply have one more example of an experiment vs. a systematic observation study.

Now, to turn to our empirical question: Did the high-drive group make more errors at the difficult choice points than did the low-drive group? To answer this question let us consider the number of errors made by each group at the easy and at the difficult choice points (Table 8.10). There we can see, for instance, that the high-drive subject who ranked highest in total number of errors made 10 errors at the easy choice points and 11 at the difficult choice points. The difference between the latter and the former is entered in the "Difference" column of Table 8.10. We can also see that the pairmate for this subject made six errors at the easy choice points and 14 at the difficult choice points, the difference being eight errors.

Table 8.10. *Number of Errors Made at the Easy and Difficult Choice Points as a Function of Drive Level.*

Level on Initial Measure	HIGH-DRIVE SUBJECTS			LOW-DRIVE SUBJECTS		
	Choice Point Easy	Choice Point Difficult	Difference	Choice Point Easy	Choice Point Difficult	Difference
1	10	11	1	6	14	8
2	8	10	2	4	14	10
3	4	9	5	2	12	10
4	4	8	4	3	9	6
5	4	7	3	4	6	2

Think about these data for a minute. If the high-drive group made more errors at the difficult choice points than did the low-drive group, then the difference scores in Table 8.10 should be greater for the high-drive group. And, to test the prediction, they should be *significantly* greater. Consequently, we need to obtain the difference between these difference scores and to compute a matched *t*-test on them. We have entered the difference scores of Table 8.10 in Table 8.11 and computed the difference between these difference scores under the column labeled "*D*."

Table 8.11. *Difference Between Number of Errors on the Easy and Difficult Choice Points as a Function of Drive Level.*

Level on Initial Measure	Difference for High-Drive Subjects	Difference for Low-Drive Subjects	<i>D</i>
1	1	8	-7
2	2	10	-8
3	5	10	-5
4	4	6	-2
5	3	2	1
$\bar{X}_{HD} = 3.00$			$\Sigma D = -21$
$\bar{X}_{LD} = 7.20$			$\Sigma D^2 = 143$

That is, the difference between the number of errors at the easy and difficult choice points for the top ranked subject of the high-drive group was one, and for the pairmate it was eight. The difference between these two values is -7 . And so on for the remaining pairs of subjects. We now seek to test the scores under " D " to see if their mean is significantly different from zero. Equation (8.1) requires the following values, computed from Table 8.11:

$$\bar{X}_{HD} = 3.00$$

$$\bar{X}_{LD} = 7.20$$

$$\Sigma D = -21$$

$$\Sigma D^2 = 143$$

$$n = 5$$

Substituting these values into Equation (8.1):⁸

$$t = \frac{7.20 - 3.00}{\sqrt{\frac{143 - \frac{(-21)^2}{5}}{5(5-1)}}} = 2.53$$

Entering our Table of t (p. 108) with a value of 2.53 and 4 df , we can see that a t of 2.776 is required for significance at the .05 level. Hence, we cannot reject the null hypothesis and thus cannot assert that variation in drive level resulted in different performance at the difficult choice points. In fact, we can even observe that the direction of the means is counter to that of the prediction, i.e., the low-drive group actually made more errors than did the high-drive group. Speculations about why the prediction was not confirmed were offered by the original authors. The interested student is encouraged to consult the articles dealt with in that publication, but our primary purpose here is accomplished by this realistic consideration of a matched groups design.

STATISTICAL ANALYSIS WITH THE A-TEST

The statistical test that we have presented for analyzing the two matched groups design has been the t -test. We may use different statistical tests in analyzing the results obtained from any particular experimental design.

⁸Incidentally, we might make use of a principle of statistics in computing the numerator of the t -test for the matched-groups design: that the difference between the means is equal to the mean of the differences of the paired observations. Therefore, as a shortcut, instead of computing the means of the two groups and subtracting them, as we have done, we could divide the sum of the differences (ΣD) by n and obtain the same answer:

$$\frac{\Sigma D}{n} = \frac{21}{5} = 4.20.$$

The t -test has been used for the matched-groups design for several reasons: (1) because we used it for the randomized groups design, it was convenient to expand it for the present design; (2) because it is the test that is generally used; and (3) because we later want to consider several important characteristics of experimentation by referring to the generalized formula for t as it

Table 8.12. *Table of A.*

For any given value of $n - 1$, the table shows the values of A corresponding to various levels of probability. A is significant at a given level if it is equal to or less than the value shown in the table.

$n - 1$	PROBABILITY					$n - 1$
	0.10	0.05	0.02	0.01	0.001	
1	0.5125	0.5031	0.50049	0.50012	0.5000012	1
2	0.412	0.369	0.347	0.340	0.334	2
3	0.385	0.324	0.286	0.272	0.254	3
4	0.376	0.304	0.257	0.238	0.211	4
5	0.372	0.293	0.240	0.218	0.184	5
6	0.370	0.286	0.230	0.205	0.167	6
7	0.369	0.281	0.222	0.196	0.155	7
8	0.368	0.278	0.217	0.190	0.146	8
9	0.368	0.276	0.213	0.185	0.139	9
10	0.368	0.274	0.210	0.181	0.134	10
11	0.368	0.273	0.207	0.178	0.130	11
12	0.368	0.271	0.205	0.176	0.126	12
13	0.368	0.270	0.204	0.174	0.124	13
14	0.368	0.270	0.202	0.172	0.121	14
15	0.368	0.269	0.201	0.170	0.119	15
16	0.368	0.268	0.200	0.169	0.117	16
17	0.368	0.268	0.199	0.168	0.116	17
18	0.368	0.267	0.198	0.167	0.114	18
19	0.368	0.267	0.197	0.166	0.113	19
20	0.368	0.266	0.197	0.165	0.112	20
21	0.368	0.266	0.196	0.165	0.111	21
22	0.368	0.266	0.196	0.164	0.110	22
23	0.368	0.266	0.195	0.163	0.109	23
24	0.368	0.265	0.195	0.163	0.108	24
25	0.368	0.265	0.194	0.162	0.108	25
26	0.368	0.265	0.194	0.162	0.107	26
27	0.368	0.265	0.193	0.161	0.107	27
28	0.368	0.265	0.193	0.161	0.106	28
29	0.368	0.264	0.193	0.161	0.106	29
30	0.368	0.264	0.193	0.160	0.105	30
40	0.368	0.263	0.191	0.158	0.102	40
60	0.369	0.262	0.189	0.155	0.099	60
120	0.369	0.261	0.187	0.153	0.095	120
∞	0.370	0.260	0.185	0.151	0.092	∞

applies to the matched-groups design. There is, however, a computationally simpler test that can be used for this design — the A -test (Sandler, 1955).^a

^aAppropriate for a two tailed test and for the usual null hypothesis that $\mu_1 - \mu_2 = 0$.

This statistic Equation (8.3) has been rigorously derived from the t ratio; hence the two tests always yield the same conclusions as far as level of significance is concerned. The equation for computing A is:

$$(8.3) \quad A = \frac{\Sigma D^2}{(\Sigma D)^2}$$

To illustrate the computation of A , we need merely refer to the data presented in Table 8.11. There we found that $\Sigma D = -21$, and $\Sigma D^2 = 143$. Hence

$$A = \frac{143}{(-21)^2} = 0.32. \text{ We now need to determine the probability level for } A.$$

To do this we compute the degrees of freedom just as we did for Equation (8.1), i.e., $df = n - 1$. Since $n = 5$ in this example, $df = 4$. And entering Table 8.1 with four degrees of freedom we read across the rows. We note that as we move to the right, the value of A , unlike the value of t in the t table, decreases. That is, the *smaller* the value of A , the smaller the probability associated with it. In the t table the *larger* the value of t , the smaller the associated value of P . Hence, we find that with $df = 4$, a value of $A = 0.32$ has a P of less than 0.10 but greater than 0.05. And this is precisely what the results of the t -test told us. For further practice with the A -test, you might check the results of the first experiment (Table 8.4) that we used as an example. (You should find that $A = 0.23$ and thus $P < 0.05$). The computational advantages of the A -test over the t -test for the two-matched-groups design should be apparent. Not only is A quicker and easier to compute than t , and thus less conducive to error, but it is also more accurate since fewer rounding errors are involved. Certainly you should prefer the A -test in a two-matched-group design.

WHICH DESIGN TO USE: RANDOMIZED GROUPS OR MATCHED GROUPS?

One advantage of the matched-groups design over the randomized-groups design is that it assures approximate equality of the two groups. That equality is not helpful to us, however, unless it is equality as far as measures of the dependent variable are concerned. Hence, if the matching variable is highly correlated with the dependent variable, then the equality of groups is beneficial. If not, then it is not beneficial — in fact, it can be detrimental. If a high correlation obtains, we should prefer a matched-groups design. But how high is high? We shall now offer some guiding considerations to help answer this question. To do this let us point out a general disadvantage of the matching design. We have said that the formula for computing degrees of freedom is $n - 1$. The formula for degrees of freedom with the randomized-groups design is $N - 2$. In other words when using the matched-groups

design you have fewer degrees of freedom available than with the randomized-groups design, assuming equal numbers of subjects in both designs. For instance, if there are seven subjects in each group, $n = 7$, or $N = 14$. With the matched-groups design we would have $7 - 1 = 6$ degrees of freedom whereas for the randomized-groups design, we would have $14 - 2 = 12$. And we may recall that the greater the number of degrees of freedom available, the smaller the value of t required for significance, other things being equal. For this reason the matched-groups design suffers a disadvantage compared to the randomized-groups design.

It may happen that a given t would have been significant with the randomized-groups design but not with the matched-groups design. Suppose we obtained the same value of t regardless of the design used. The t , we might say, is 2.05 obtained with 16 subjects per group. With a matched-groups design we would have 15 df and find that a t of 2.131 is required for significance at the 5 per cent level — hence the t is not significant; but with the 30 df available we would need a value of only 2.042 for significance at the 5 per cent level.

To summarize the situation concerning the choice of a matched-groups or a randomized-groups design, the advantage of the former is that the value of t may be increased if there is a positive correlation between the matching variable and the dependent variable. On the other hand, one loses degrees of freedom when using the matched-groups design; half as many degrees of freedom are available with it as with the randomized-groups design. Therefore, if the correlation is going to be large enough to more than offset the loss of degrees of freedom, then one should use the matched-groups design.¹⁰ If it is not, then the randomized-groups design should be used.¹¹ In short, if one is to use the matched groups design he should be rather sure that the correlation between his matching and his dependent variable is rather high and positive.

At this point a bright student might say: "Look here, you have made so much about this correlation between the matching and the dependent variable, and I understand the problem. You say to try to find some previous evidence that a high correlation exists. But maybe this correlation doesn't hold up in your own experiment. I think I've got this thing licked. Let's match our subjects on what we think is a good variable and then actually compute the correlation. If we find that the correlation is not sufficiently high, then let's forget that we matched subjects and simply run a t -test for a

¹⁰It might be observed that if the number of subjects in a group is large (e.g., if $n = 30$), then one can afford to lose degrees of freedom by matching. That is, there is such a small difference between the value of t required for significance at any given level with a large df that one would not lose much by matching subjects even if the correlation between the independent and dependent variables is zero. Hence the loss of df consideration is only an argument against the matched groups design when n is relatively small.

¹¹An elaboration of these statements is offered in the Appendix.

randomized-groups design. If we do this, we can't lose; either the correlation is pretty high and we offset our loss of degrees of freedom using the matched-groups design or it is too low so we use a randomized-groups design and don't lose our degrees of freedom."

"This student," we might say, "is thinking, and that's good. But what he's thinking is wrong." An extended discussion of what is wrong with the thinking must be left to a course in statistics, but we can say that the error is similar to that previously referred to in setting the level of significance for t (p. 109). There we said that the experimenter may set whatever level of significance he desires, providing he does it before he conducts his experiment. Analogously, the experimenter may select whatever design he wishes, providing he does it before he conducts his experiment. And in either case the decision must be adhered to. For where he chooses a matched-groups design he has also mortgaged himself to a certain type of statistical test (e.g., the matched t -test, which has a certain probability attached to its results). And if he changes the design he disturbs the probability that he can assign to his t through the use of the t table. Hence, if an experimenter decides to use a matched-groups design, he must stick to that decision. Perhaps the following experience might be consoling to you in case you ever find yourself in the unlikely situation described. The author once helped conduct an experiment in which a matched-groups design was used (McGuigan & MacCaslin, 1955a). Previous research had shown that the correlation between the variable that was used to match subjects and the dependent variable was 0.72. This was an excellent opportunity to use a matched-groups design. However it turned out that the correlation was -0.24 for the data collected. And we shall see in the appendix what a negative correlation does to the value of t . Hence in the author's experiment not only were degrees of freedom lost, but (the value of the statistical test that would correspond to) the value of t was actually decreased.

APPENDIX TO CHAPTER 8

It is particularly advantageous to consider several additional matters that should facilitate an understanding of the two-matched-groups design. To do this we shall consider the *generalized* equation for the t -test, *generalized* in that it is applicable to either the randomized-groups or the matched-groups design. It may be written:¹²

$$(8.4) \quad t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - 2(r_{12})\left(\frac{s_1}{\sqrt{n_1}}\right)\left(\frac{s_2}{\sqrt{n_2}}\right)}}$$

¹²It is important not to forget that sample characteristics (statistics) are used as estimates of corresponding population characteristics (parameters). The (population) parameters

The two previous equations for t [Equations (5.2) and (8.1)] are derivatives of Equation (8.4). To understand Equation (8.4) we need to understand the following new symbols: s (the standard deviation) and s^2 (the variance), both of which are measures of variability, and r_{12} .

THE STANDARD DEVIATION AND VARIANCE

Suppose that we are interested in making some statements about the intelligence of the students at a given college. There may be 1000 students in the college; this would give us 1000 scores with which to deal, a very cumbersome number. If someone asked us about the intelligence of the students in the college we might start reading the scores to him. Before we could reach the thousandth score, however, our inquirer undoubtedly would have withdrawn his question. A much more reasonable procedure for telling him about the intelligence scores of our college students would be to resort to certain summary statements about them. We could, for instance, compute a mean and tell our inquirer that the mean intelligence of the student body is 125, or whatever. Although this would be accurate, it would not be adequate, for there is more to the story than that. Whenever we seek to describe a group of data we need to offer two kinds of statistics — *a measure of central tendency and a measure of variability*. Measures of central tendency tell us something about the central point value of a group of data. They are kinds of averages that tell us what the typical score in a distribution of data is. The most common measure of central tendency is the mean. Measures of variability, tell us how the scores are spread out; they indicate something about the nature of the distribution of scores. In addition to telling us this, they also tell us about the range of scores in the group. The most frequently used measure of variability, probably because it is usually the most reliable of these measures (in the sense that it varies least from sample to sample), is the standard deviation. The standard deviation is symbolized by s .

To illustrate the importance of measures of variability we might imagine that our inquirer says to us: "Fine. You have told me the mean intelligence of your student body, but how homogeneous are your students? Do their scores tend to concentrate around the mean, or are there many that are

are fixed, but unknown, while the (sample) statistics can be expected to vary from sample to sample. Since our emphasis is on how to compute statistics, we are using notations for sample statistics in our discussions. Your further work will lead to a greater appreciation of the distinction, which can be more clearly made by contrasting the notation for statistics and parameters, such as \bar{X} and μ for the mean, s and σ for the standard deviation, and r and ρ for correlation.

considerably below the mean?" To answer this, we might resort to the computation of the standard deviation. The larger the standard deviation, the more variable are our scores. To illustrate, let us assume that we have collected the intelligence scores of students at two different colleges. Plotting the number of people who obtained each score at each college we might obtain the distributions shown in Figure 8.1.

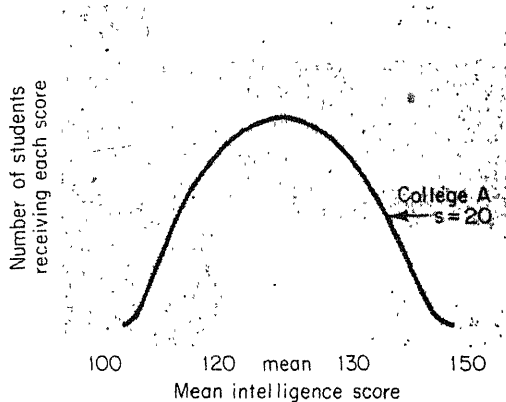


FIGURE 8.1.

Distribution of intelligence scores at two colleges.

By computing the standard deviation¹³ for the two groups, we might find their values to be 20 for College A and 5 for College B. Comparing the

¹³Note again that we are primarily concerned with values for samples. From the sample values the population values may be inferred. This is another case where we must limit our consideration of statistical matters to those that are immediately relevant to the conduct of experiments. But you are again advised to pursue these important topics by further work in statistics. An equation for computing the standard deviation for

a sample is $s = \sqrt{\frac{n\sum X^2 - (\sum X)^2}{n(n-1)}}$. Or, if you have already computed the SS , the equation

is $s = \sqrt{\frac{SS}{n-1}}$. As Guilford (1965) points out, the estimate of the population value

for the standard deviation would be slightly different but when N is large (30 or greater) the two computed values are practically identical. You should be able to compute s on the basis of the knowledge that you now possess. To check yourself you might compute s for the dependent variable scores of the two groups in Table 8.3. You should find that for the Reading group $s = 2.61$ and for the Reading and Reciting group $s = 2.17$.

Incidentally, with regard to the second equation above it is worth your while to examine Equation (5.2) on p. 101. There you can see that the denominator of the t ratio contains the components for computing s . In computing t you can, therefore, quickly compute s for your two groups. They are valuable to have in studying and reporting your data and, furthermore, the two values of s^2 can be easily compared to see if you have satisfied the assumption of homogeneity of variance (cf., p. 354).

distributions for the two colleges, we note that there is considerably more variability in College A than in College B. That is, the scores for College A are more spread out or scattered than for College B. And this is precisely what our standard deviation tells us; the larger the value for the standard deviation, the greater the variability of the distribution of scores. The standard deviation (for a normal distribution) also gives us the more precise bit of information that about two-thirds of the scores fall within the interval that is one standard deviation above and one standard deviation below the mean. To illustrate, let us first note that the mean intelligence of the students of the two colleges is the same, 125. If we subtract one standard deviation (i.e., 20) from the mean for College A and add one standard deviation to that mean, we obtain two values: 105 ($125 - 20 = 105$) and 145 ($125 + 20 = 145$). Therefore about two-thirds of the students in College A have an intelligence score between 105 and 145. Similarly, about two-thirds of the students at College B have scores between 120 ($125 - 5$) and 130 ($125 + 5$). Hence, we have a further illustration that the scores at College A are more spread out than at College B. We might for a moment speculate about these student bodies. College A, we might guess, is rather lenient in its selection of students, as might be the case in some state universities. College B is much more selective, having a rather homogeneous student body. Such a sample might occur for a private institution with high tuition costs. In any event we wish to make only one point here, that the larger the value of the standard deviation, the more variable (spread out) the scores.

The symbol s^2 is known as the *variance* of a set of values. It has essentially the same characteristics as the standard deviation and is merely the square of the standard deviation. Hence, if $s = 5$, then $s^2 = 25$. To illustrate these statistics further, let us assume that we have obtained the dependent variable scores for the two groups in an experiment shown in Table 8.13.

Table 8.13. *Some Dependent Variables Scores for Two Groups of Subjects.*

EXPERIMENTAL GROUP Subject	GROUP Score	CONTROL GROUP Subject	GROUP Score
1	10	1	7
2	1	2	6
3	0	3	7
4	5	4	5
5	3	5	6
6	7	6	7
7	9	7	6
8	6	8	5
9	8	9	7
10	2	10	7

The scores for the experimental group vary from 0 to 10; the standard deviation here is 3.48. The scores for the control group, on the other hand, are much less variable. In fact, all the scores are either five, six, or seven. We should expect that the standard deviation for the control group is considerably smaller than for the experimental group. We shall not be disappointed, for its computation yields a value of .82. In turn, the variance is 12.10 for the experimental group and .68 for the control group. Using the variances as indices of variability, we can see that they also show that the variability of the experimental group is greater than that of the control group. Incidentally, we might note that if all the scores for one group were the same, say seven, both the standard deviation and the variance would be zero, for there would be zero variability among the scores.

THE NATURE OF r_{12}

With this understanding of the standard deviation and variance, let us now turn to the symbol r_{12} , the last unfamiliar symbol in Equation (8.4). We have already discussed the general nature of a correlation, but it remains for us to specify this particular correlation. It stands for the (linear) correlation between the dependent variable scores of the two groups of subjects in a matched-groups design. (Let us observe that r_{12} is read "the correlation between the dependent variable scores of Group 1 and Group 2" and not " r -twelve.") That is, in this type of design, we have paired subjects in one group with subjects in the other group. And these pairs are ranked on the basis of their matching variable scores. Thus, the highest pair of scores on the matching variable is ranked first, the second highest pair ranked second, and so on. If the dependent variable scores of the first pair of subjects are the highest scores in each group, if the second pair of subjects provided the second highest dependent variable scores in their respective groups, and so on down the rank of pairs without exception, then the correlation between these two sets of dependent variable scores would be perfect. That is, r_{12} would equal 1.0.¹⁴ Similarly, if there are only a few exceptions in this order, r_{12} would be high but less than perfect. And so on for the other possibilities, as discussed in the previous section on correlation.

To cement our understanding of the nature of r_{12} refer to Table 8.3 and note the sets of dependent variable scores for each of our two groups. If we correlated these two sets of scores, we would find that $r_{12} = 0.90$.¹⁵ That is, the highest pair of subjects on the matching variable has the highest set of

¹⁴Again on the assumption that the increase in scores for each group is proportional (see footnote 1, p. 167). If they are not, then the correlation will be somewhat less than 1.0.

¹⁵Here is a good example of what we said in footnote 1, i.e., although there is no exception in the rankings of the two sets of scores, the intervals between the scores in each set are not proportional, and hence the value of r only approaches 1.0.

dependent variable scores in its group, the lowest pair of subjects has the lowest dependent variable scores, and so on.

The nature of r_{12} should now be clear. And you could compute it for any set of data that you want. But what is its significance? To answer this question, let us restate a principle that was stressed earlier. For the matched-groups design to be successful, you should have a reasonably high correlation between the matching variable and the dependent variable scores. Strictly speaking, this latter correlation does not enter into your statistical analysis, but is taken account of only indirectly. That is, rather than using an actual value for the matching variable-dependent variable correlation in our t -test, we use the value of r_{12} . And this is possible because the value of r_{12} is an indication of the value of the matching variable-dependent variable correlation.¹⁶ That is, if the matching variable-dependent variable correlation is high, r_{12} will be high; and if the matching variable-dependent variable correlation is low, r_{12} will also be low. The reasonableness of these statements should be apparent after a little reflection. The only reason that we would expect a high value of the correlation between the paired dependent variable scores of our two groups is that they were matched together on the basis of a variable that correlates with the dependent variable. Thus, the reason that the pair of subjects who were ranked first on the matching variable should both exhibit the highest scores on the dependent variable is that there is a correlation between the matching variable and the dependent variable. On the other hand, if the correlation between the matching and the dependent variable scores is zero, we would expect the value of r_{12} to be zero, for there would be no reason to expect that the top-ranked pair of subjects on the matching variable should both exhibit the highest dependent variable scores.

The situation at this point is that r_{12} is the correlation between the dependent variable scores of pairs of subjects in a matched-groups design. Since r_{12} is an indication of the value of the correlation between the matching and the dependent variable scores, we use it in conducting our statistical analysis.

Let us now take a broader look at Equation (8.4). If we want to increase our chances of obtaining a significant t there are two courses of action that can be followed. First, we can attempt to increase the value of the numerator (the difference between the means of the two groups), or to decrease the value of the denominator. As the value of the numerator increases, or as the value of the denominator decreases, the value of t increases. And we know that the larger t is, the more likely it is to be significant. To illustrate, let us say that the numerator is five and the denominator is ten. In this case t is:

$$t = \frac{5}{10} = 0.50$$

¹⁶Again, emphasizing that we are considering only linear relationships, ignoring curvilinearity.

But if we are able to decrease the value of the denominator to two, with no change in the numerator, we would have:

$$t = \frac{5}{2} = 2.50$$

And a t of 2.50 is likely to be significant, whereas one of 0.50 is not.

Our question should now be how we can decrease the value of the denominator of Equation (8.4). This matter is discussed more thoroughly in Chapter Fifteen, but let us consider one possible way here. We may note that the larger the variance of the two groups, the larger is the denominator. If s_1^2 and s_2^2 are each ten, the denominator will be larger than if they are both five. But we may note that from the variances we subtract r_{12} (and also s_1 and s_2 , but these need not concern us here). And any subtraction from the variances of the two groups will result in a smaller denominator with, as we said, an attendant increase in t . Furthermore, we said, the size of r_{12} depends on the correlation between the matching variable and the dependent variable. Hence, if that correlation is large and positive, we may note that the denominator is decreased.

By way of illustration, assume that the difference between the means of the two groups is 5 and that there are 9 subjects in each group (n_1 and n_2 both equal 9). Further assume that s_1 and s_2 are both 3 (hence s_1^2 and s_2^2 are both 9) and that r_{12} is 0.70. Substituting these values in Equation (8.4) we obtain:

$$t = \frac{5}{\sqrt{\frac{9}{9} + \frac{9}{9} - 2(0.70)\left(\frac{3}{\sqrt{9}}\right)\left(\frac{3}{\sqrt{9}}\right)}} = 6.49$$

It should now be apparent that the larger the positive value of r_{12} , the larger is the term that is subtracted from the variances of the two groups. In an extreme case of the above illustration, where $r_{12} = 1.0$, we may note that we would subtract 2.00 from the sum of the variances (2.00); this leaves a denominator of zero, in which case t might be considered to be infinitely large. On the other hand, suppose that r_{12} is rather small — say it is 0.10. In this case we would merely subtract 0.20 from 2.00, and the denominator would be only slightly reduced. Or if $r_{12} = 0$, then it can be seen that zero would be subtracted from the variances, not reducing them at all. The lesson should now be clear: *the larger the value of r_{12} (and hence the larger the value of the correlation between the matching variable and the dependent variable), the larger the value of t .*

Now let us consider the heart of the matter: How large should be the correlation between the matching and the dependent variable in order to prefer a matched-groups design to a randomized-groups design? Well, we can't answer this question, at least not in this form. But we can give a very good answer by changing it somewhat. This is, in fact, the principal reason why

we have stressed the nature of r_{12} and its relationship with the correlation between the matching and the dependent variable. Thus, we shall answer the question: How large should r_{12} be before a matched-groups design should be preferred? And the answer is given in Fig. 8.2.¹⁷ To use Figure 8.2 you

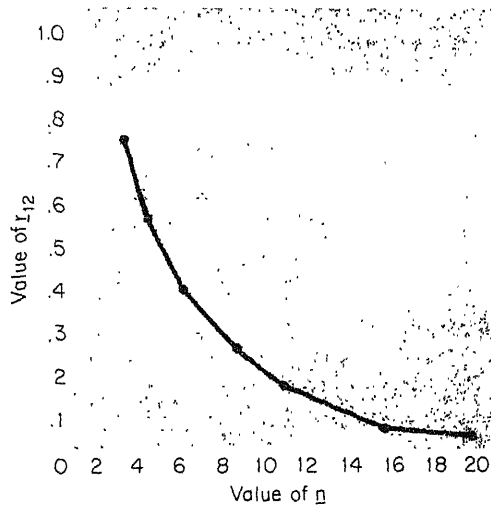


FIGURE 8.2.

A relationship between n and r_{12} . Enter with a value of n and read the expected value of r_{12} that intersects the curve at that point. If your expected value of r_{12} exceeds the value obtained from the curve, a matched-groups design is to be preferred.

merely select the number of subjects in each group (n) and enter the figure at that value on the horizontal axis. Then read up until you reach the curve. The value of the curve at that point on the vertical axis indicates the necessary minimum value of r_{12} in order for a matched-groups design to be preferred. For instance, if you want to know the minimal value of r_{12} that you will need in the event that you have 12 subjects in each group, you would find "12" on the horizontal axis. Then reading up to the curve and over to the vertical axis you find that r_{12} is about 0.17. *In order to prefer a matched-groups design, then, you should have a minimal value of 0.17 for r_{12} .*

One final consideration of the value of r_{12} is what the effect of a negative correlation would be on the value of t . A little reflection should reveal that a negative correlation *increases* the denominator, thus decreasing t . In this case, instead of subtracting from the variances, we would have to add to them ("a minus times a minus gives us a plus"). Furthermore, the larger

¹⁷We are simplifying the situation by presenting a rather conservative case. It can rather safely be said that if the value of your r_{12} is at least that indicated for a certain number of subjects in Fig. 8.2, then you should prefer a matched-groups design.

the negative correlation, the larger our denominator becomes. For example, suppose that in the previous example instead of having a value of $r_{12}=0.70$, we had $r_{12} = -0.70$. In this case we can see that our computed value of t would decrease from 6.49 to 2.72. That is,

$$t = \frac{5}{\sqrt{\frac{9}{9} + \frac{9}{9} - 2(-0.70)\left(\frac{3}{\sqrt{9}}\right)\left(\frac{3}{\sqrt{9}}\right)}} = 2.72$$

We previously said that Equation (8.4) is a generalized formula, applicable to either of the two designs that we have discussed. One might ask, however, in what way it is applicable to the randomized-groups design, for it contains a correlation term and we have not referred to any correlation when using it; it is absurd, for instance, to talk about the correlation between pairs of subjects on the dependent variable when using the randomized-groups design, for by its very nature subjects are not paired. The answer to this is that since subjects have not been paired the correlation between any random pairing of subjects in the long run is zero. That is, if we randomly selected any subject in an experimental group, paired him with a randomly selected subject in the control group, and continued this procedure for all subjects, we would expect the correlation between the dependent variable scores to be zero (or more precisely, the correlation would not be significantly different from zero). There simply would be no reason to expect other than a zero correlation since the subjects were not paired together on more than a chance basis. When using the randomized-groups design, we assume that r_{12} of Equation (8.4) is zero. And being zero, the term that includes r_{12} "drops out." Thus, Equation (8.4) assumes the following form for the randomized-groups design:

$$(8.5) \quad t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

One final note. We have labeled the type of design discussed in this chapter as the *matched-groups design* where we have limited our discussion to the case of two groups. The two groups may be said to be matched because we paired subjects with similar scores. Since all subjects were paired together, the groups had to be approximately equivalent. This fact may be determined by comparing the distribution of matching scores for the two groups. The best such comparison would probably be to compare the means and standard deviations of the two groups. We would expect to find that the two groups would be quite similar on these two measures. However, the same result could also be achieved in other ways. That is, two groups could be formed in other ways so that they would have similar means and standard deviations.

For example, we could simply assign subjects to two groups so that their total scores would be similar; no subjects would be paired together, but the means and standard deviations of the two groups would be approximately the same. Therefore, it may be considered that the technique of pairing subjects together is a specific type of design that results in matched groups. For this reason it could as well be called the *paired-groups design* to distinguish it from alternative procedures that result in matched groups. Alternative procedures, however, require different statistical analyses from those presented here (McNemar, 1962). Since they are not so generally used, nor judged to be as effective, they will not be considered further.

The two-matched-groups design (or if you prefer, the two-paired-groups design) implies that the design could be extended to more than two groups. For a discussion of a matched-groups design for more than two groups you are referred to Edwards (1968).

SUMMARY OF THE COMPUTATION OF t FOR A TWO-MATCHED-GROUPS DESIGN

Assume that two groups of subjects have been matched on an initial measure as indicated, and that the following dependent variable scores have been obtained for them.

<i>Initial Measure</i>	<i>Group 1</i>	<i>Group 2</i>
1	10	11
2	10	8
3	8	6
4	7	7
5	7	6
6	6	5
7	4	3

1. The equation for computing t , Equation (8.1), is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{n}}{n(n-1)}}$$

2. Compute the value of D for each pair of subjects, and then the sum of D ($\sum D$), the sum of the squares of D ($\sum D^2$), the sum of D squared $[(\sum D)^2]$, and n .

<i>Initial Measure</i>	<i>Group 1</i>	<i>Group 2</i>	<i>D</i>
1	10	11	-1
2	10	8	2
3	8	6	2
4	7	7	0
5	7	6	1
6	6	5	1
7	4	3	1
			$\Sigma D = 6$
			$\Sigma D^2 = 12$
			$n = 7$

3. Determine the difference between the means. This may be done by computing the mean of the differences. Since the latter is easier, we shall do this.

$$\text{Mean of the differences} = \frac{\Sigma D}{n} = \frac{6}{7} = 0.86$$

4. Substitute the above values in Equation (8.1):

$$t = \frac{0.86}{\sqrt{\frac{12 - \frac{(6)^2}{7}}{7(7-1)}}$$

5. Perform the operations as indicated and determine the value of t .

$$t = \frac{0.86}{\sqrt{0.16}} = 2.15$$

6. Determine the number of degrees of freedom associated with the computed value of t .

$$df = n - 1 = 7 - 1 = 6$$

7. Enter the table of t with the computed values of t and df . Determine the probability associated with this value of t . In this example, $0.1 > P > 0.05$. Therefore, assuming a significance level of 0.05, the null hypothesis is not rejected.

PROBLEMS

1. A psychologist seeks to test the hypothesis that the western grip for holding a tennis racket is superior to the eastern grip. He matches his subjects on the basis of a physical fitness test, trains them in the use of these two grips, respec-

tively, and obtains the following scores on their tennis-playing proficiency. Assuming adequate controls, that a 0.05 level of significance is set, and that the higher the score the better the performance, what can he conclude with respect to his empirical hypothesis?

<i>Rank on Matching Variables</i>	<i>Score on Dependent Variable</i>	
	<i>Eastern Grip Group</i>	<i>Western Grip Group</i>
1	10	2
2	5	8
3	9	3
4	5	1
5	0	3
6	8	1
7	7	0
8	9	1

2. To test the hypothesis that the higher the induced anxiety, the better the learning, an experimenter formed two groups of subjects by matching them on an initial measure of anxiety. Next he induced considerable anxiety into his experimental group, but not into his control group. He then obtained the following scores on a learning task, where the higher the score, the better the learning. Assuming adequate controls were exercised and that he set a significance level of 0.05, did he confirm his hypothesis?

<i>Rank on Matching Variable</i>	<i>Dependent Variable Scores</i>	
	<i>Experimental Group</i>	<i>Control Group</i>
1	8	6
2	8	7
3	7	4
4	6	5
5	5	3
6	3	1
7	1	2

3. A military psychologist wishes to evaluate a training aid that was designed to facilitate the teaching of soldiers to read a map. He forms two groups of subjects, matching them on the basis of a visual perception test (an ability that is important in the reading of maps). He sets a significance level of 0.02 and exercises proper controls. Assuming that the higher the score, the better the performance, did the training aid facilitate map reading proficiency?

<i>Rank on Matching Variable</i>	<i>Scores of Group That Used the Training Aid</i>	<i>Scores of Group That Did Not Use the Training Aid</i>
1	30	24
2	30	28
3	28	26
4	29	30
5	26	20
6	22	19
7	25	22
8	20	19
9	18	14
10	16	12
11	15	13
12	14	10
13	14	11
14	13	13
15	10	6
16	10	7
17	9	5
18	9	9
19	10	6
20	8	3

EXPERIMENTAL DESIGN

The Case of More Than Two Randomized Groups

CONCERNING THE VALUE OF USING MORE THAN TWO GROUPS

Among the more frequently used designs in psychological research are those that employ more than two groups. Suppose a psychologist had two methods of remedial reading available. The methods are both presumably helpful to students who have not adequately learned to read by the usual method, but he wishes to know which method is superior. Furthermore, he wants to know whether either of these methods is actually superior to the normal method for such problem cases. To answer these questions, he might design an experiment that involves three groups of subjects.

If he has available 60 students who show marked reading deficiencies, his first step would be randomly to assign them to three groups. Assume that he assigns an equal number of subjects to each group, although this need not be the case. The first group would be taught to read by using Method A and the second group by Method B. A comparison of the results from these two

groups would tell him which, if either, is the superior method. He also wants to know if either method is superior to the normal method of teaching, which has heretofore been ineffective with this group. So, he would have his third group continue training under the normal method, as a control group. After a certain period of time, perhaps nine months, a standard reading test might be administered to the three groups. A comparison of the reading proficiency of the three groups on this test should answer the questions.

It is also possible to answer these questions by conducting a series of separate two-groups experiments. It would be possible, for instance, to conduct one experiment in which Method A is compared to Method B, a second in which Method A is compared to the control condition, and a third experiment in which Method B is compared to the control condition. Such a procedure is obviously less desirable, for not only would more work be required but the problem of controlling extraneous variables would be sizeable. For example, we would wish to hold the experimenter variable constant, so the same experimenter should conduct all three experiments. Even so, it is likely that the experimenter would behave differently in the first and last experiments, perhaps due to improvement in his teaching proficiency, or even because of boredom or fatigue. Therefore, the design in which three groups are used simultaneously is superior in that less work is required, fewer subjects are used, and experimental control is better.

The randomized-groups design for the case of more than two groups may be applied to a wide variety of problems. To illustrate we might list some problems suggested by Edwards that are amenable to this type of design: "... the influence of different periods of food deprivation upon learning; the influence of different numbers of reinforcements upon conditioning and extinction; the influence of different methods of instructions upon achievement; the influence of different sets of verbal instructions upon problem solving; the influence of different sensory cues upon maze learning; the influence of different kinds of motivation upon performance; the influence of different periods of rest upon fatigue; the influence of different kinds of interpolated activities upon learning; the influence of different kinds of work situations upon production and fatigue; the influence of different periods of practice upon learning" (Edwards, 1950, p. 185).

The procedure for applying a multi-group design (i.e., a design with more than two groups) to any of the above problems would be to select several values of the independent variable and assign a group of subjects to each value. For example, to study the influence of different periods of food deprivation upon performance, we might choose the following values of the independent variable: 0 hours, 1 hour, 12 hours, 24 hours, 36 hours, and 48 hours of deprivation. Having selected six values of the independent variable, we would have six different groups of subjects, probably animals. To study

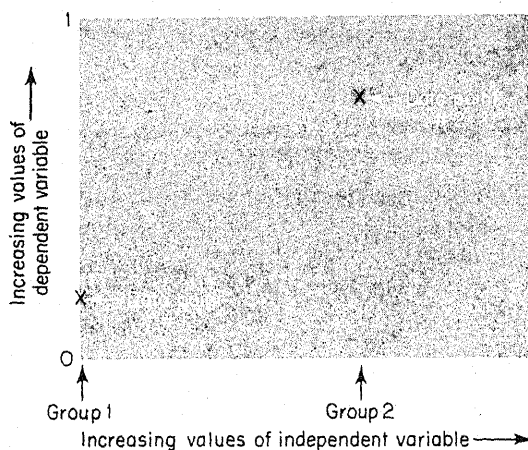
the influence of different periods of practice upon learning we might select four values of the independent variable: 0, 5, 10, and 15 hours. We would then randomly assign our subjects to four groups and train one group under each condition.

These considerations now make apparent yet another advantage of a multi-groups over a two-groups design. That is that if you attempted to attack any of the above problems by means of a two-groups design, you would have to decide which two of many values of the independent variable to employ. A sizeable advantage of a multi-groups design becomes apparent when one tries to decide which two values of the independent variables to use. Consider the above example concerning the influence of different periods of practice upon performance. We previously selected four values of the independent variable to study. Which two would we use for a two-groups design? It would be advisable to have a control condition, so we would probably choose a zero hour period for one group. The second group might be trained under a five-hour period.

Now, let us imagine that the four-groups design yields the following results: no difference in performance among the zero, 5, and 10 hour conditions, but the 15 hour condition is superior to the first three. The conclusion from this four-group experiment would be that variation of the length of practice from zero to 10 hours does not affect performance, however, greater periods of practice increases proficiency. But if the two-groups design (using only 0 and 5 hour practice periods) were applied to this problem, the results would *suggest* that variation of the length of practice does not affect performance, a conclusion that would be in error. Thus, in general, it should be apparent that the more values of the independent variable sampled, the better our evaluation of its influence on a given dependent variable.

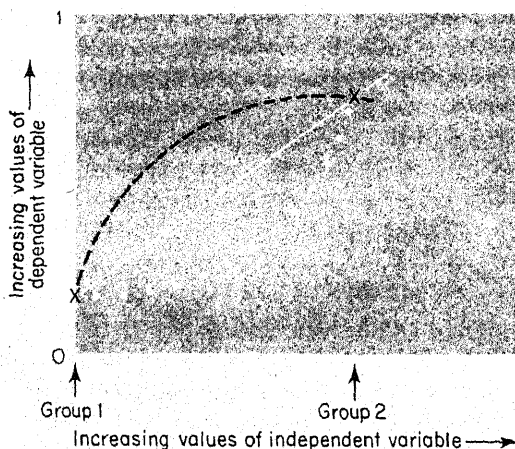
Research in any given area usually progresses through two stages: first, we seek to determine which of many possible independent variables influences a given dependent variable; and second, when a certain independent variable has been identified as influential on a dependent variable, we attempt to establish the precise relationship between them. Even though it is possible that a two-groups design would accomplish the first purpose, it could not accomplish the second. For an adequate relationship cannot usually be established with only two values of the independent variable (and therefore also only two values of the dependent variable). To illustrate this point refer to Figure 9.1, where the values of an independent variable are indicated on the horizontal axis, and the corresponding values of the dependent variable are read on the vertical axis.¹ The two plotted points (obtained from a two-groups design) indicate that as the value of the independent variable increases,

¹The range of the independent variable in the following discussions should be clear from the context, e.g., from zero to infinity. We shall also assume that the data points are highly reliable and thus not the product of random variation.

**FIGURE 9.1.**

Two data points obtained from a two-groups design. Group 1 was given a zero value of the independent variable while Group 2 was given a positive value. The value of the dependent variable is less for Group 1 (data point #1) than for Group 2 (data point #2).

the mean value of the dependent variable also increases. But this is a crude picture, for it tells us nothing about what happens between (or beyond) the two plotted points. See Figure 9.2 to illustrate a few of the infinite number of possibilities.

**FIGURE 9.2.**

The actual relationship between the independent and the dependent variable is partially established by the two data points. However, the curves that may pass through the two points are infinite in number.

By using a three-groups design the relationship may be established more precisely. Let us say that we have the same two groups as in Figure 9.1, but in addition we have a third group that received a value of the independent variable halfway between those of the other two groups. Assuming that the mean dependent variable value for Group 3 is that depicted in Figure 9.3,

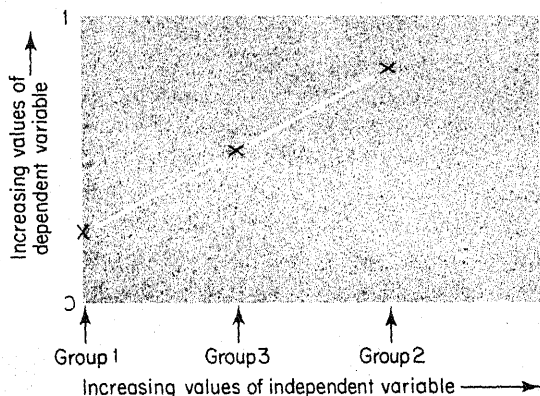


FIGURE 9.3.

The addition of a third data point (Group 3) suggests that the relationship is a linear function.

we would conclude that the relationship is probably a linear (straight-line) function. Of course, we might be wrong. That is, the relationship is not necessarily the straight line indicated in Figure 9.3, for it is possible that some other relationship is actually the "true" one, such as one of those shown in Figure 9.4. Nevertheless, with only three data points we prefer to bet that

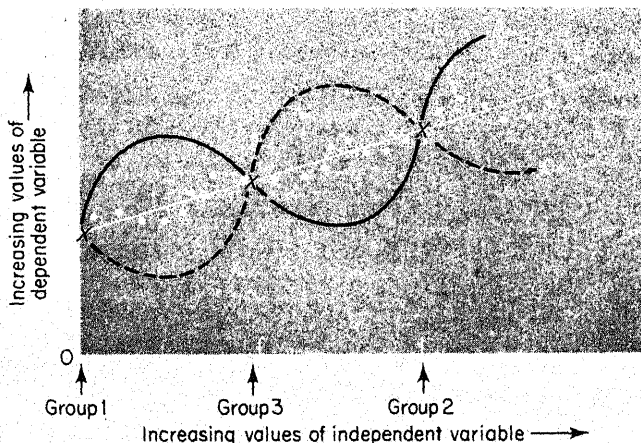


FIGURE 9.4.

Other curves may possibly pass through the three data points.

the straight line is the "true" relationship because it is the simplest of the several possible relationships. And experience suggests that it is reasonable to assume that the simplest curve gives the best predictions. That is, we would predict that if we obtain a data point for a new value of the independent variable (in addition to the three already indicated in Figure 9.3) the new data point would fall on the straight line. Different predictions would be made from the other curves of Figure 9.4.

To illustrate, suppose that we add a fourth group whose independent variable value is halfway between those of Groups 1 and 3. On the basis of the four relationships depicted in Figure 9.4 we could make four different predictions about the dependent variable value of this fourth group. First, using the straight-line function, we would predict that the data point for the fourth group would be that indicated by X_1 in Figure 9.5, i.e., if the straight

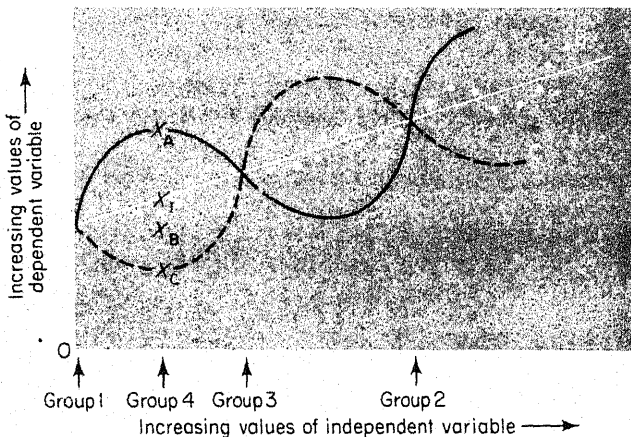


FIGURE 9.5.

Four predictions of a data point for Group 4. From the straight line of Figure 9.3, we would predict that the dependent variable score would be indicated by X_1 . From the three curves of Figure 9.4 (curves A, B, and C), we would predict that the data point would be that indicated by the X_A , the X_B , and the X_C respectively.

line is the "true" relationship, the data point for the fourth group should fall on that line. The three curves of Figure 9.4, however, lead to three additional (and different) predictions.

Assume that when the data for the fourth group are analyzed, they indicate that the mean score is actually that indicated by the X_1 of Figure 9.5. This increases our confidence in the straight-line function; it, rather than the other possible functions, is probably the "true" one. If these were actually the results, our procedure of preferring the simplest curve as the "true" one (at least until contrary results are obtained) is justified. This procedure is

what Reichenbach (1938) calls *inductive simplicity*, the selection of the simplest curve that fits the data points. The safest induction is that the simplest curve provides the best prediction of additional data points. With the randomized design for more than two groups you can establish as many points as you like, consistent with the effort you can expend.

One general principle of experimentation when using a two-groups design is that it is advisable to choose rather extreme values of the independent variable.² If we had followed this principle, we would not have erred in the example concerning the influence of the period of practice upon performance. For instead of choosing 0 and 5 hour periods, as we did, we perhaps should have selected zero- and 15-hour periods. In this event the two-groups design would have led to a conclusion more in line with that of the four-groups design. However, it should still be apparent that the four-groups design yielded considerably more information, allowing us to establish the relationship between the two variables with a high degree of confidence. Even so, the selection of extreme values for two groups can lead to difficulties in addition to those already considered. To illustrate, assume that two data points are obtained, such as those indicated by the X's in Figure 9.6. Our conclu-

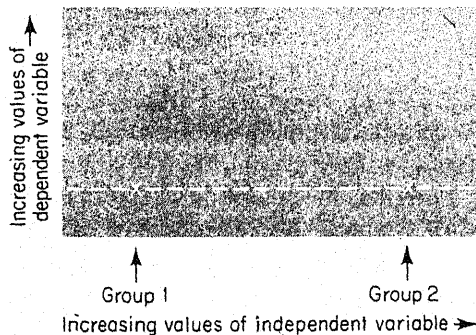


FIGURE 9.6.

Two data points for extreme values of the independent variable using a two-groups design. These points suggest that the independent variable does not affect the dependent variable.

sion would probably be that manipulation of the independent variable does not influence the dependent variable, for the dependent variable values for the two groups are the same. The best guess is that there is a lack of relation-

²Let us emphasize the word "rather," for seldom would we want to select independent variable values for two groups that are really extreme. This is so because it is likely that all generalizations in psychology break down when the independent variable values are unrealistically extreme. Weber's law, which you probably studied in introductory psychology, is a good example. For although Weber's law holds rather well for weights that you can conveniently lift, it would obviously be absurd to state that it is true for extreme values of weights such as those of atomic size or those of several tons.

ship as indicated by the horizontal straight line fitted to the two points.³ Yet the actual relationship may be that indicated in Figure 9.7, a relation-

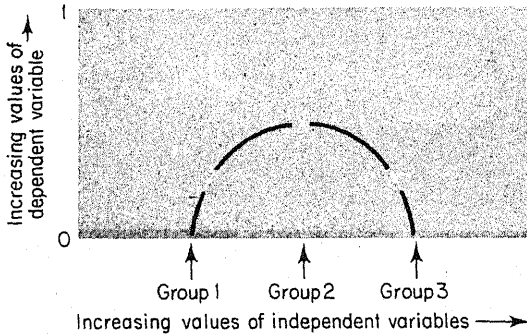


FIGURE 9.7.

Postulated actual relationship for the data points of Figure 9.6. This relationship would be uncovered by a suitable three-groups design.

ship that probably would have been uncovered by the use of a three-groups design. The corresponding principle with a three-groups design would be to select two rather extreme values of the independent variable and also one value midway between them. Of course if the data point for Group 3 had been the same value as for Groups 1 and 2, then we would be more confident that the independent variable did not affect the dependent variable.

To summarize what we have said we may note that psychologists seek to determine which of a number of independent variables influence a given dependent variable and also attempt to establish the relationship between them. With a two-groups design one is never sufficiently sure that he has selected the appropriate values of the independent variable in his attempt to determine whether or not that variable is effective. By using more than two groups, however, he increases his chances of: (1) accurately determining whether a given independent variable is effective; and (2) specifying the relationship between the independent and the dependent variable.⁴

With these principles behind us, let us now be more concrete with regard to the potential pitfalls in using a two-groups design. To illustrate, consider an experiment by Bersh (1951) on the development of secondary reinforcement. The general principle that accounts for the development of this

³One might quibble and say that the horizontal line is still a relationship. However, from our point of view this "relationship," if true, would indicate that variation of the independent variable does not affect the dependent variable.

⁴Data from Edgington (1964) suggests a decrease in the frequency of use of two groups designs, e.g., in 1948, 51 per cent of psychological articles sampled employed the *t*-test, but only 19 per cent did so in 1962. Multi-groups designs are being increasingly used, perhaps for reasons cited above.

important phenomenon states that any neutral stimulus (such as a light) that is associated with a primary reinforcer (such as food) will itself acquire the power of acting as a reinforcer. To determine that a stimulus has become a secondary reinforcer, one can present it after an organism makes a response to see if that response increases in strength. For instance, after a light has been associated with food (and presumably thereby acquired secondary reinforcing properties) an experimenter could place a rat in a Skinner Box that contains a bar that can be pressed. Then, whenever the rat presses the bar the light is presented. If the bar-pressing response increases in frequency it may be concluded that the light has acquired the status of a secondary reinforcer.

Bersh's problem was set by the failure of Schoenfeld, Antonitis, and Bersh (1950) to establish a light as a secondary reinforcer, even though they *had* associated the light with food. The reason for the failure, they suggested, was because of an ineffective time interval between the onset of the light and the presentation of the food. Bersh's problem; therefore, was to ascertain the effect of varying this temporal interval; his independent variable was the length of time that the light was on prior to the delivery of the pellet. His procedure was to place a rat in a Skinner Box with the bar temporarily removed. He then presented a light to the animal and, after the lapse of some specific amount of time, he delivered a pellet of food. Once the light and food had been associated a number of times, the bar was replaced in the Skinner Box and the animal was allowed to press it. Each depression of the bar resulted in the onset of the light. The dependent variable was the number of bar pressing responses that occurred within a ten minute period, so that the greater the number of responses, the stronger the secondary reinforcing properties of the light.

Now, place yourself in the position of the experimenter as he designed this experiment. In the training phase you present a light to the rat, after which you deliver a pellet of food. If you use a two-groups design, what two time values would you select to separate these two presentations? As more or less of a control condition you would probably want to use a zero value, i.e., you would probably present the light and food simultaneously with no time intervening. But what would be the value for your second condition? Suppose that, because you had to do something,⁵ you decided to turn on the light one second before the delivery of the food.

If you actually conducted this experiment your results should resemble those reported by Bersh, i.e., the animals who had a zero second delay be-

⁵In research, as in many phases of life, one frequently faces problems for which no appropriate response is available. A principle that the author has found useful was given by a college mathematics teacher (Dr. Bell) to be applied when confronted with an apparently unsolvable math problem: "If you can't do anything, do something." You will be delightfully amazed at the frequency with which this principle leads, if not directly, at least indirectly, to success.

tween light onset and delivery of food would make approximately 19 bar presses within the ten minute test period. But approximately 25 responses would be made by the animals for whom light preceded food by one second during training. Hence, the light acquires stronger secondary reinforcing properties when it precedes food by one second than when it occurs simultaneously with food. May it now be concluded that the longer the time interval between presentation of light and food, the stronger the reinforcing properties of the light? To study this question we have plotted the number of responses made for these two conditions in Figure 9.8, and have fitted a straight line to them. Before we can have confidence in this conclusion, we must face gnawing questions such as what would have happened had there been a 0.5 second delay or a two-second delay. Would dependent variable values for these conditions have fallen on the straight line, as suggested by the two circles in Figure 9.8? The answer, of course, is that we would never

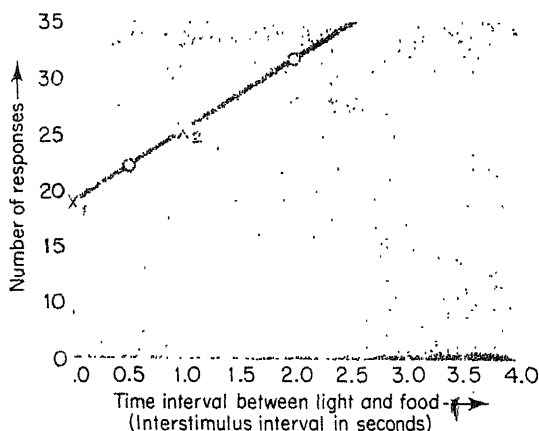


FIGURE 9.8.

Two data points for a two-groups design. Data point #1 (indicated by X_1) resulted from a zero second time interval during secondary reinforcement training, and data point #2 (X_2) resulted from a one-second delay. The suggestion is that the longer the time interval, the larger the number of resulting responses. Hence the prediction for other time interval values, such as for 0.5 and 2.0 seconds, are indicated by the circles.

know unless there were an experiment involving such conditions. Fortunately, in this instance, relevant data are available, in addition to the 0.0 second and the 1.0 second delay conditions, Bersh also ran groups of rats with delays of 0.5 seconds, 2.0 seconds, 4.0 seconds and 10.0 seconds, and the complete curve is presented in Figure 9.9. By studying Figure 9.9 we can see how erroneous would be the conclusion based on the two-groups experiment. Instead of a 0.5 second delay resulting in about twenty-two responses, as was predicted by Figure 9.8, this value led to results about the same as a

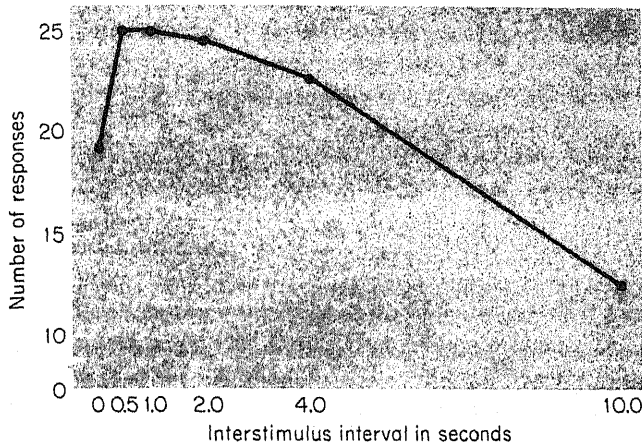


FIGURE 9.9.

Strength of secondary reinforcement as a function of the interstimulus interval during training. The measures are the number of responses during the first ten minutes of the test session (Bersh, 1951).

1.0 second delay, after which the curve, instead of continuing to rise, falls rather dramatically. In short, then, the conduct of a two-groups experiment on this problem would have, in all likelihood, resulted in an erroneous conclusion. The truth is that strength of secondary reinforcement increases from a zero to a 0.5 second delay, is approximately the same for a 0.5 to a 1.0 second delay, after which it decreases. And this complex relationship could not possibly have been determined by means of a single two-groups design. Thus, in general, it should be apparent that the more values of the independent variable sampled, the better our estimation of its influence on a given dependent variable.

STATISTICAL ANALYSIS OF A RANDOMIZED-GROUPS DESIGN WITH MORE THAN TWO GROUPS

As in previous designs, we need to determine whether our groups differ significantly. However, we now have several groups to compare. As before we shall compare groups by comparing differences among their means. But what statistical procedure is most appropriate for this type of problem? Unfortunately for our present purposes, there is much disagreement among statisticians and among psychologists as to the correct answer to this question. In part, though only in part, the disagreements stem from different types of hypotheses that are being tested, and different aspects of the question that are

emphasized. We here wish to minimize the extent to which we enter the various controversies, and to this end shall say that we are interested in doing the following: (1) in making comparisons only between pairs of individual groups, that is, we are not interested in combining two or more groups to test these combined groups against some other group or combination of groups; and (2) in making comparisons between all possible combinations of the separate groups taken two at a time. To emphasize these points, let us say that you conduct an experiment in which there are three groups. In this event, we are saying that you would be primarily interested in determining whether Group 1 differs from Group 2, whether Group 1 differs from Group 3, and whether Group 2 differs from Group 3. You would not, for example, be interested in determining whether Groups 1 and 2, considered together as one group, differ from Group 3. With this limitation of our interest (although we shall comment briefly on procedures where our interest might be otherwise), we shall consider two types of statistical analysis for the more-than-two-randomized-groups design. After prolonged investigation of the numerous statistical procedures available, it was the author's opinion that Duncan's Range Test is the most appropriate. However an alternative procedure that is often used by psychologists starts with the analysis of variance. By presenting both of these procedures (with the author's recommendation of Duncan's Range Test) you can make your own decision as to which you prefer to apply. It might be added, incidentally, that Duncan's Range Test is easier to apply, and is less time consuming.⁶

DUNCAN'S RANGE TEST: A THREE-GROUPS DESIGN

To apply Duncan's Range Test let us consider an experiment by Sulzbacher (1967) on programmed learning. The standard method of studying a programmed textbook is for the student to read each frame and to write in the missing words. For example, a frame may be:

"Underwood is an experimental psychologist,
Lewis is an experimental psychologist and
Duncan is an _____."

Experimental
Psychologist

Immediately on writing in the appropriate words the student then receives confirmation of his response by uncovering the correct answer, here placed at the right.

Sulzbacher's question concerned the value of writing the answer in each frame as against just thinking the answer without actually writing it in. Ac-

⁶This recommendation from the first edition has received confirmation with the increasing use of Duncan's Range test within the past years. *

cordingly he formed three groups of elementary school pupils and had each study a mathematics program. The first group took the program in the normal manner and thus was called the Overt Responding (OR) Group. The second group just "thought" the answers without writing them [The Covert Responding (CR) Group] and then received confirmation. A control group merely read through the program with the answers already filled

Table 9.1. *G Ratios for Three Methods of Studying a Programmed Text.*

Overt Responders (OR)		Covert Responders (CR)		Non-Responders (NR)	
.46	.75	.26	.42	.51	.83
.36	.87	.57	.81	.78	.45
.56	.24	.82	.40	.65	.53
.28	.29	.94	.78	.64	.55
.72	.50	.85	.89	.33	.05
.80	.41	.57	.43	.65	.62
.59	.87	.00	.32	.58	.57
.60	.48	.90	.56	.50	.58
.91	.78	.62	.58	.56	.48
.93	.86	.70	.71	.38	.68
$n = 20$		$n = 20$		$n = 20$	
$\Sigma X = 12.26$		$\Sigma X = 12.13$		$\Sigma X = 10.92$	
$\Sigma X^2 = 8.5072$		$\Sigma X^2 = 8.5147$		$\Sigma X^2 = 6.4922$	
$\bar{X} = .613$		$\bar{X} = .606$		$\bar{X} = .546$	

in for them, and the response confirmation portion was blacked out [Non-Responders (NR) group]. A *G* ratio was computed for each student, and the resulting scores are presented in Table 9.1.⁷

We seek to test for significant differences among the three groups or, more precisely, among the means of the three groups. The first step is to compute the sum of squares (*SS*) of the dependent variable scores for each group. Recall that the equation for the sum of squares of any group is:

$$(9.1) \quad SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

⁷*G* is a ratio of amount actually learned to amount that could possibly have been learned. The higher the *G* value, the greater the amount learned from the program. *G* is computed as follows: The mean score on the pretest is calculated. After the students study, the test is again administered and the mean post test score is computed. The difference between the pre and the post test scores indicates the *actual gain*. The difference between the pretest score and the possible score is a measure of *possible gain*. By dividing the possible gain into the actual gain one can tell how much the students learned relative to how much they could have learned, as measured by the test. The advantage of *G* is that it takes into account the amount of knowledge that the students had prior to their study of the experimental material. A straight gain score does not do this. For instance, if the students score 90 per cent on the pretest, the actual gain score may be artificially restricted. *G* may well find a wider application than just in the evaluation of programmed texts (cf. McGuigan & Peters, 1965).

Hence it can be seen that we need ΣX^2 , ΣX , and n for each group. These values are computed as before and are presented in Table 9.1. We merely need to substitute them separately for each group in Equation (9.1). For Group OR the SS is:

$$SS_{OR} = 8.5072 - \frac{(12.26)^2}{20} = .9918$$

The SS of Groups CR and NR can be seen to be:

$$SS_{CR} = 8.5147 - \frac{(12.13)^2}{20} = 1.1579$$

$$SS_{NR} = 6.4922 - \frac{(10.92)^2}{20} = .5299$$

Next we compute the square root of the error variance (s_e) (see p. 356), which for three groups, is given by the equation:

$$(9.2) \quad s_e = \sqrt{\frac{SS_A + SS_B + SS_C}{3(n-1)}}$$

Where n is still the number of subjects in each group. Substituting the sum of squares for the three groups and n in Equation (9.2) we find that s_e is:

$$s_e = \sqrt{\frac{.9918 + 1.1579 + .5299}{3(20-1)}} = .2168$$

We now determine the degrees of freedom for s_e appropriate for Duncan's Range Test, which is given by the equation:

$$(9.3) \quad df = N - r$$

N is the total number of subjects in the experiment, and r is the number of groups. This is found to be:

$$df = 60 - 3 = 57$$

Assuming that we have set a 5 per cent level of significance for testing the difference between two means, we now need to refer to Table 9.2.⁸ There we see that the columns are labeled "Number of Groups" and the rows " df ." The values in the table are the "least significant standardized ranges," which are symbolized r_p .⁹ We shall require various values of r_p for each test between two means that we shall make. Since we have three means in the

⁸See Duncan (1955) for an elaboration of the precise nature of this probability value.

⁹Do not confuse r_p with r , which is the number of groups. It is unfortunate that we cannot take the time to lay a broad base for our techniques of statistical analysis and for discussing these concepts in greater detail. The interested reader can of course always correct this deficiency by referring to other sources. For an elaboration of the concepts in this chapter you might refer to Duncan (1955), Li (1957), or to the more psychologically oriented works of Edwards (1968), Hays (1963), Ray (1960), and Winer (1962).

present example we shall make the following tests: between the extreme means of three groups; between the highest and the middle means; and, between the lowest and the middle means. The p that is the subscript to the r indicates these three situations. For the first, $p = 3$ and hence $r_p = r_3$, and for the other two $p = 2$ and hence $r_p = r_2$. These symbols may be written r_3 and r_2 . Therefore, we need to enter Table 9.2 to obtain the value of r_p when the extreme value of a group of three means is being tested, which is the column labeled "3." Finding the column labeled "3," we read down rows until we come to the appropriate degrees of freedom. However, 57 df is not in the table. The best that we can do is to find a row for 40 df and for 60 df . The values of r_p for three groups (i.e., r_3) for 40 and 60 df are 3.01 and 2.98, respectively. By interpolating (linearly) we find that the desired value of r_3 is 2.98. This value shall be used indirectly in comparing the groups with the highest and the lowest means. More specifically, it is used for comparing the extreme means of a group of three means. But we will also need to compare adjacent groups. Since adjacent groups involve the comparison of only two means we need to follow the same procedure for the column labeled 2; i.e., for two groups. Reading down column 2 until we come to 40 and 60 df we find, by interpolating, that r_2 is 2.83 for 57 df . These two values of r_p are written in the top row of Table 9.3. If the procedure for selecting and using the values for r_p is not clear, you should proceed through this section, accepting our statements on faith, for we discuss the matter more thoroughly later.

Table 9.3. Values of r_p and R_p for 2 and 3 Groups with 57 df .

	Number of Groups	
	2	3
r_p	2.83	2.98
R_p	.137	.145

Our next step is to compute the "least significant ranges" for our values, which is symbolized by R_p , where:

$$(9.4) \quad R_p = s_e r_p \sqrt{\frac{1}{n}}$$

Therefore, to find R_p for each number of groups we need to multiply our computed value of s_e by the appropriate value of r_p and $\sqrt{1/n}$. Since our computed value of $s_e = .217$ (rounded off), r_p for two groups is 2.83, and $n = 20$, R_2 is:

$$R_2 = (.217) (2.83) \sqrt{\frac{1}{20}} = .137$$

Similarly R_3 is:

$$R_3 = (.217) (2.98) \sqrt{\frac{1}{20}} = .145$$

These values of R_p have been entered in the bottom row of Table 9.3. Now let us order (rank) the means for our three groups from the lowest to the highest. They were found to be .613, .606, and .546 for Groups OR, CR, and NR respectively. Hence, our ordering is:

	<i>Group</i>		
	<i>NR</i>	<i>CR</i>	<i>OR</i>
Mean:	.546	.606	.613

Now, the final step is to compare the differences between our ordered means and the values of R_p . Starting with highest and the lowest means it can be seen that the difference between Groups OR and NR is .067. By comparing Group OR with Group NR we have compared the extreme means of three groups. Therefore, we read the value of $R_3 = .145$ from Table 9.3. This means that the difference between the group with the highest mean (Group OR) and the group with the lowest mean (Group NR) must exceed .145 for that difference to be significant at the five per cent level. We found that the difference between Group OR and Group NR was .067. Since this value does not exceed .145 we may conclude that the sample mean of Group OR is not significantly different from the sample mean of Group NR; had, on the other hand, the mean difference between these two groups been .150, say, they would have been significantly different.

Should we now determine whether Group OR is different from Group CR? Not in this example, but to illustrate for other cases, the procedure would be as follows. Since these means are adjacent in our ranked order, we are comparing two means. Hence, we read from Table 9.3 that the appropriate value of $R_2 = .137$. The obtained difference between the means of these two groups is .007. If the obtained difference exceeds the value of .137, these two groups differ beyond the .05 level of significance. Since .007 does not exceed .137, our conclusion is in the negative; the sample mean of Group OR is not significantly different from the sample mean of Group CR.

At this point we may note a rule of general procedure. Since the difference between the extreme means (Groups OR and NR) was not significant, there was no necessity for testing the lesser difference between Groups OR and CR. For if there is no significant difference between the most extreme means (here between Groups OR and CR), any lesser difference can not possibly be significant.

As another example, the remaining possible comparison would be between Groups CR and NR. The difference between their means is .060. Since

we would now be comparing two groups, we would read the R_p value of .137 from Table 9.3. As before, the necessary difference between a group of two means must exceed .137. Since .060 does not exceed .137, we conclude that Groups CR and NR do not differ significantly, a fact that we already knew. On the other hand, if the extreme means do differ significantly, the only way to find out whether the lesser differences are significant is to proceed with the test. This is the reason that we illustrated the procedure even though it was inappropriate for this case.

The major finding of this study, thus, is that the three groups did not differ significantly, when considered pairwise, on the criterion measure. Consequently variation of modes of responding failed to influence amount learned.

The general rule for selecting and using the various values of r_p is: After the group means have been ordered, count the number of groups *between* the two that you are testing and add that number to those two. Then, enter the table of r_p (Table 9.2) for that number of groups. In the three-group design just discussed we have the following situation:

<i>Group NR</i>	<i>Group CR</i>	<i>Group OR</i>
\bar{X}_1	\bar{X}_2	\bar{X}_3

If we are comparing Groups OR and NR, we count one group between these two, giving us a total of three. Hence, we enter column 3 of Table 9.2. A test between Groups OR and CR, however, has no groups intervening between them. Therefore, we add zero to our count of two, indicating that we enter column 2 of Table 9.2. And likewise, when we are comparing Groups CR and NR, the number is two. Let's say that we have four groups and are testing the highest mean against the lowest. In this case, two groups intervene. Therefore, we add two to the groups that we are testing, indicating that we should enter Table 9.2 at the column labeled 4. However, we will also need to compare the highest group with the second lowest. In this case, one group will intervene, making our count three. The next comparison will be between the two highest groups, a situation in which no groups intervene. Hence, the count would be two, and we should record that value of r_p for further use. Similarly, if we have a five groups design and are testing the highest mean against the lowest, the count would be five. For testing the highest against the middle group, the count would be three. And so forth.

A FIVE GROUPS DESIGN

The statistical analysis for Duncan's Range Test should now be clear. The same procedure is followed for experiments involving more than three groups

with several minor computational differences. To illustrate its application to a design involving five groups we shall select from the data presented by Woods and Holland (1966). The particular aspect of their study on which we shall focus concerns the relative preference of rats for various water temperatures.

More specifically, each of five groups of rats was presented with a choice of two temperatures of water, and each animal had free access to either temperature. Group 1 had a choice of either 24 degree Centigrade water or 26 degree water. Group 2 could spend their time in water of either 28 or 30 degrees temperature. Similarly, the choices for Groups 3, 4, and 5 were 30 or 32 degrees, 32 or 34 degrees, and 36 or 38 degrees, respectively.¹⁰ The question was which of the two choices the animals in each group would prefer. The percentage of time spent in the warmer of the two temperatures is presented for each animal in Table 9.4.¹¹

Table 9.4. *Percentage of Time Spent in Warmer Temperature (Groups are Identified by Degrees of Warmer Side).*

1(26°)	2(30°)	Group 3(32°)	4(34°)	5(38°)
61	69	62	27	17
74	67	46	36	23
64	50	57	50	15
59	53	64	49	20
69	61	79	50	17
80	56	49	36	15
77	64	47	27	14
70	72	54	33	30
19	46	40	33	22
69	73	53	63	28
<hr/>				
$\Sigma X: 642$	611	551	404	201
$\Sigma X^2: 43,886$	38,141	31,481	17,598	4,321
$n: 10$	10	10	10	10
$\bar{X}: 64.2$	61.1	55.1	40.4	20.1

¹⁰A more powerful, but more complex, method of statistical analysis may be used when the independent variable (e.g., number of trials) is quantified along some dimension, e.g., a regression analysis, such as the method of orthogonal polynomials, as in Cochran and Cox (1957). This does not mean, however, that Duncan's Range Test is inappropriate for this type of situation.

¹¹The values reported here are the result of an arc sin transformation of the raw data, but that need not detain us.

The values of ΣX , ΣX^2 , n , and \bar{X} have been computed for each group. To compute the sum of squares for each group we merely substitute the values required by Equation (9.1):

$$SS_1 = 43,886 - \frac{(642)^2}{10} = 2669.6$$

$$SS_2 = 38,141 - \frac{(611)^2}{10} = 808.9$$

$$SS_3 = 31,481 - \frac{(551)^2}{10} = 1120.9$$

$$SS_4 = 17,598 - \frac{(404)^2}{10} = 1276.4$$

$$SS_5 = 4,321 - \frac{(201)^2}{10} = 280.9$$

The previous formula for s_e was specialized for the case of three groups. An equation applicable to any number of groups is:

$$(9.5) \quad s_e = \sqrt{\frac{SS_1 + SS_2 + SS_3 + \cdots + SS_r}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_r - 1)}}$$

The numerator of this general equation for computing s_e merely indicates that you should add the sums of squares for all of the groups together, where of course r indicates the number of groups. The denominator tells you to add the number of subjects in each group minus one together, continuing for all r groups. Since we have five groups in the present example, we have five sums of squares to add, and since $n_1 = n_2 = n_3 = n_4 = n_5$ we may merely multiply $(n - 1)$ by 5. Hence Equation (9.5) may be written for this case as:

$$(9.6) \quad s_e = \sqrt{\frac{SS_1 + SS_2 + SS_3 + SS_4 + SS_5}{5(n - 1)}}$$

Substituting the sum of squares and n for the five groups in Equation (9.6) we obtain:

$$s_e = \sqrt{\frac{2669.6 + 808.9 + 1120.9 + 1276.4 + 280.9}{5(10 - 1)}} = 11.70$$

The appropriate number of degrees of freedom is:

$$df = 50 - 5 = 45$$

For a 5 per cent level test we enter Table 9.2 to obtain the necessary values of r_p . We have five groups so we enter the column labeled 5. Reading down to 40 and 60 df we find (by interpolating) that the value of r_5 for 45 df is 3.16. Similarly the value of $r_4 = 3.10$, or $r_3 = 3.00$ and of $r_2 = 2.85$. These values are entered in the top row of Table 9.5.

Table 9.5. *Values of r_p and R_p for Five Groups with 45 df.*

	2	3	4	5
r_p	2.85	3.00	3.10	3.16
R_p	10.5	11.1	11.5	11.7

To obtain the R_p values we multiply the appropriate value of s_e by r_p and $\sqrt{1/n}$ (Equation (9.4)). These computations are:

$$\text{For 2 groups: } R_p = (11.70) (2.85) \sqrt{1/10} = 10.5$$

$$\text{For 3 groups: } R_p = (11.70) (3.00) \sqrt{1/10} = 11.1$$

$$\text{For 4 groups: } R_p = (11.70) (3.10) \sqrt{1/10} = 11.5$$

$$\text{For 5 groups: } R_p = (11.70) (3.16) \sqrt{1/10} = 11.7$$

These values have been entered in the bottom row of Table 9.5.

We next order our means from lowest to highest:

	<i>Group</i>				
	5	4	3	2	1
Mean	20.1	40.4	55.1	61.1	64.2

Following our previous procedure we first must test the difference between the extreme means of five groups. The extreme means are for Groups 1 and 5, their difference being 44.1. Since we are considering five groups, we read the value of $R_5 = 11.7$ from Table 9.5 (for five groups: three intervening groups, plus the two we are testing). The difference of 44.1 is larger than 11.7 and, therefore, we may conclude that the mean of Group 1 is significantly larger than the mean of Group 5 at the 5 per cent level. The next comparison is between Groups 1 and 4. The difference between their means is 23.8. Now we are comparing the difference between extreme values of four groups. Hence, we read R_p from Table 9.5 as 11.5. Since 23.8 exceeds 11.5, we may conclude that the mean of Group 1 is significantly larger than the mean of Group 4. The next comparison is between Group 1 and Group 3. Three means enter our consideration. Therefore the appropriate R_p from Table 9.5 is 11.1. The difference in means between Groups 1 and 3 is 9.1. Since this difference does not exceed 11.1 the means of these two groups are not significantly different. From this we also can immediately conclude that the means of Groups 2 and 1, as well as the means of Groups 3 and 2 are not significantly different. In short, we have found that Group 1 is significantly different from Groups 5 and 4, but is not significantly different from Groups 3 and 2; furthermore, that Groups 2 and 3 are not significantly different. This fact is indicated by the following scheme. Any two means that are

underscored by the same line are *not significantly different*. Any two means that are *not underscored* by the same line are *significantly different*. Memorize these two sentences. Since Group 1 is not significantly different from Groups 2 and 3, we draw a line under those three means. But since Group 1 is significantly different from Groups 4 and 5, we do not extend the line under them. That is:

	Group				
	5	4	3	2	1
Mean	20.1	40.4	<u>55.1</u>	61.1	<u>64.2</u>

We have now tested for significant differences between Group 1 and the other groups. Our next step is to compare Group 2 with Groups 4, and 5. The extreme values among these four groups occur for Groups 2 and 5, their mean difference being 40.0. From Table 9.5 we find that the R_p for comparing four groups is 11.5. Since 40.0 is larger than the necessary value of 11.5 we may say that Groups 2 and 5 differ significantly. Now let us move on to Group 4. The mean difference between Groups 2 and 4 is 20.07. From Table 9.5 we find that the appropriate value of R_p is 11.1; 20.7 exceeds 11.1. Therefore, Groups 2 and 4 differ significantly. A summary of our findings to this point remains as it was, i.e.,

	Group				
	5	4	3	2	1
Mean	20.1	40.4	<u>55.1</u>	61.1	<u>64.2</u>

We now proceed to test for significant differences between Group 3 and Groups 4 and 5. The difference between the means of Groups 3 and 5 is 35.0. From Table 9.5 we find that the value of R_p for comparing three groups is 11.1. Our obtained difference exceeds this value. We, therefore, conclude that Group 3 is significantly different from Group 5. To compare Groups 3 and 4 we find that their mean difference is 14.7. R_p for comparing two groups is 10.5. Therefore, these two groups differ significantly. These findings are indicated by not drawing a common line under Groups 3, 4, and 5, and hence all lines remain as they were:

	Group				
	5	4	3	2	1
Mean	20.1	40.4	<u>55.1</u>	61.1	<u>64.2</u>

Our final comparison is between Groups 4 and 5. The difference between their means may be seen to be 20.3. From Table 9.5 we find a value of 10.5

for R_p when two groups are being tested. Since 20.3 exceeds 10.5, we conclude that there is a significant difference between Groups 4 and 5. This finding is indicated by not drawing a line under Groups 4 and 5, so the schema does not change.

	Group				
	5	4	3	2	1
Mean	20.1	40.4	55.1	61.1	64.2

The above line (and lack of lines) under the means of the groups constitute a summary of our findings. The line under Groups 1, 2, and 3 indicates that these groups are not significantly different. But since that line does not extend to the other groups, we know that the means of Groups 1, 2, and 3 are significantly higher than for Groups 4 and 5. The lack of a common line under Groups 4 and 5 indicates that the mean of Group 4 is significantly higher than the mean of Group 5. In short, of the ten pairwise comparisons, only three pairs are not significantly different, i.e., based on pairwise comparisons we conclude that:

1. Groups 1, 2, and 3 do not differ significantly.
2. Groups 1, 2, and 3 have significantly higher means than do Groups 4 and 5.
3. Group 4 has a significantly higher mean than does Group 5.

In Figure 9.10 we have plotted the relationship between the independent and the dependent variables. Recall that the rats had a choice of either a warmer or a colder temperature. The mean percentage of the time that they

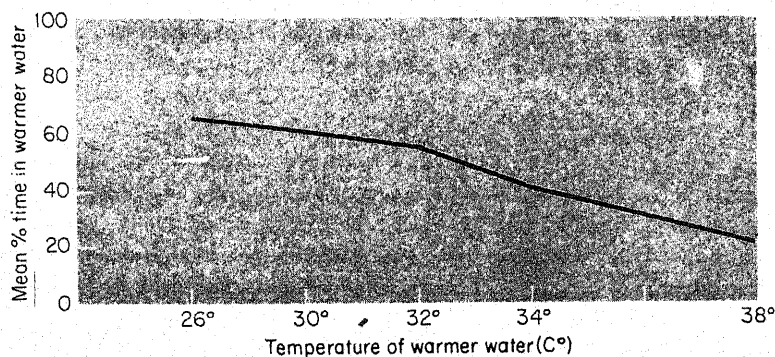


FIGURE 9.10.

Preference for the warmer of two choices of water as water temperature increases. As temperature increases, preference for warmer water decreases. Values on the vertical axis are actually arc sin transformations of the raw data.

chose the warmer temperature is plotted along the vertical axis, and the temperature of the warmer water is plotted along the horizontal axis. The warmer water was, for each choice, two degrees Centigrade above that of the colder water. For example, Group 1 had a choice of 24 vs. 26 degree water, and their mean time in the 26 degree water was 64.2 percent, a value that is plotted above the 26 degree point on the horizontal axis. Similarly, Group 2 spent 61.1 per cent of their time in the 30 degree as against the 28 degree water. Following along the horizontal axis, we can see that the amount of time spent in the warmer water continues to decrease until Group 5 chose the 38 degree water only 20.1 per cent of the time as against 79.9 per cent for the colder water. In short, as temperature of water increased, the per cent of time that the animals chose the warmer water decreased. The most interesting aspect of these data, incidentally, occurs for the 32 and 34 degree points on the horizontal axis. Skin temperature of the animals is, it may be noted, approximately 32 degrees. When they had to choose between 30 and 32 degrees, they preferred the 32 degree water; when they had to choose between the 32 and the 34 degree water, they still preferred the 32 degree water. Hence, rats are able to make very fine discriminations between temperatures of water, and they prefer a temperature that most closely approximates that of their skin. The reason for the results of the statistical test now become apparent. That is, when the colder water was below skin temperature, the animals preferred warmer water; hence Groups 1, 2, and 3 all went to the warmer side and were not significantly different. But when the warmer water was above skin temperature, the animals preferred the colder side, (closer to their skin temperature), and thus Groups 4 and 5 behaved significantly differently from the other groups.

In the preceding examples we have used a 5 per cent level test. Table 9.6 presents the necessary values for testing the differences between groups at the 1 per cent level. The procedure for using Table 9.6 is precisely the same as that for Table 9.2.

In order to insure that the scheme for indicating significant differences between group means is clear, let us consider several additional examples. For instance, suppose we have three groups and find that Group 3 is not significantly different from Group 2, but has a significantly higher mean than does Group 1. However, we find that there is no significant difference between Groups 1 and 2. These findings would be indicated as follows:

	Group (Increasing Means →)		
	1	2	3
Mean:	\bar{X}_1	\bar{X}_2	\bar{X}_3
	<hr/>		

Taking another example with three groups, assume that Group 3 is signifi-

Table 9.6. Values of r_p for Duncan's Range Test (Significance level = 1 per cent).

df	NUMBER OF GROUPS															
	2	3	4	5	6	7	8	9	10	12	14	16	18	20	50	100
1	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0
2	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0
3	8.26	8.5	8.6	8.7	8.8	8.9	8.9	9.0	9.0	9.0	9.1	9.2	9.3	9.3	9.3	9.3
4	6.51	6.8	6.9	7.0	7.1	7.1	7.2	7.2	7.3	7.3	7.4	7.4	7.5	7.5	7.5	7.5
5	5.70	5.96	6.11	6.18	6.26	6.33	6.40	6.41	6.5	6.6	6.6	6.7	6.7	6.8	6.8	6.8
6	5.24	5.51	5.65	5.73	5.81	5.88	5.95	6.00	6.0	6.1	6.2	6.2	6.3	6.3	6.3	6.3
7	4.95	5.22	5.37	5.45	5.53	5.61	5.69	5.73	5.8	5.8	5.9	5.9	6.0	6.0	6.0	6.0
8	4.74	5.00	5.14	5.23	5.32	5.40	5.47	5.51	5.5	5.6	5.7	5.7	5.8	5.8	5.8	5.8
9	4.60	4.86	4.99	5.08	5.17	5.25	5.32	5.36	5.4	5.5	5.5	5.6	5.7	5.7	5.7	5.7
10	4.48	4.73	4.88	4.96	5.06	5.13	5.20	5.24	5.28	5.36	5.42	5.48	5.54	5.55	5.55	5.55
11	4.39	4.63	4.77	4.86	4.94	5.01	5.06	5.12	5.15	5.24	5.28	5.34	5.38	5.39	5.39	5.39
12	4.32	4.55	4.68	4.76	4.84	4.92	4.96	5.02	5.07	5.13	5.17	5.22	5.24	5.26	5.26	5.26
13	4.26	4.48	4.62	4.69	4.74	4.84	4.88	4.94	4.98	5.04	5.08	5.13	5.14	5.15	5.15	5.15
14	4.21	4.42	4.55	4.63	4.70	4.78	4.83	4.87	4.91	4.96	5.00	5.04	5.06	5.07	5.07	5.07
15	4.17	4.37	4.50	4.58	4.64	4.72	4.77	4.81	4.84	4.90	4.94	4.97	4.99	5.00	5.00	5.00
16	4.13	4.34	4.45	4.54	4.60	4.67	4.72	4.76	4.79	4.84	4.88	4.91	4.93	4.94	4.94	4.94
17	4.10	4.30	4.41	4.50	4.56	4.63	4.68	4.72	4.75	4.80	4.83	4.86	4.88	4.89	4.89	4.89
18	4.07	4.27	4.38	4.46	4.53	4.59	4.64	4.68	4.71	4.76	4.79	4.82	4.84	4.85	4.85	4.85
19	4.05	4.24	4.35	4.43	4.50	4.56	4.61	4.64	4.67	4.72	4.76	4.79	4.81	4.82	4.82	4.82
20	4.02	4.22	4.33	4.40	4.47	4.53	4.58	4.61	4.65	4.69	4.73	4.76	4.78	4.79	4.79	4.79
22	3.99	4.17	4.28	4.36	4.42	4.48	4.53	4.57	4.60	4.65	4.68	4.71	4.74	4.75	4.75	4.75
24	3.96	4.14	4.24	4.33	4.39	4.44	4.49	4.53	4.57	4.62	4.64	4.67	4.70	4.72	4.74	4.74
26	3.93	4.11	4.21	4.30	4.36	4.41	4.46	4.50	4.53	4.58	4.62	4.65	4.67	4.69	4.73	4.73
28	3.91	4.08	4.18	4.28	4.34	4.39	4.43	4.47	4.51	4.56	4.60	4.62	4.65	4.67	4.72	4.72
30	3.89	4.06	4.16	4.22	4.32	4.36	4.41	4.45	4.48	4.54	4.58	4.61	4.63	4.65	4.71	4.71
40	3.82	3.99	4.10	4.17	4.24	4.30	4.34	4.37	4.41	4.46	4.51	4.54	4.57	4.59	4.69	4.69
60	3.76	3.92	4.03	4.12	4.17	4.23	4.27	4.31	4.34	4.39	4.44	4.47	4.50	4.53	4.66	4.66
100	3.71	3.86	3.98	4.06	4.11	4.17	4.21	4.25	4.29	4.35	4.38	4.42	4.45	4.48	4.64	4.65
∞	3.64	3.80	3.90	3.98	4.04	4.09	4.14	4.17	4.20	4.26	4.31	4.34	4.38	4.41	4.60	4.63

cantly higher than Groups 2 and 1, but there is no significant difference between Groups 2 and 1:

<i>Group</i>		
1	2	3
\bar{X}_1	\bar{X}_2	\bar{X}_3

Consider a four groups design. Assume that Group 4 is significantly superior to the other groups; that Group 3 is not significantly superior to Group 2, but is significantly superior to Group 1; and that there is no difference between Groups 1 and 2:

	<i>Group</i>			
	1	2	3	4
Mean:	\bar{X}_1	\bar{X}_2	\bar{X}_3	\bar{X}_4

With four groups again, assume that Group 4 does not differ significantly from Groups 3 and 2, but is significantly different from Group 1. Furthermore, Groups 3 and 2 are not significantly different, but Group 3 is significantly superior to Group 1. And Group 2 is significantly different from Group 1:

	<i>Group</i>			
	1	2	3	4
Mean:	\bar{X}_1	\bar{X}_2	\bar{X}_3	\bar{X}_4

STATISTICAL ANALYSIS FOR UNEQUAL *n*'s

The preceding discussion has assumed that the number of subjects in each of the several groups is equal. However, it is frequently the case that different numbers of subjects are assigned to the groups. Kramer (1956) has extended Duncan's Range Test so that it is applicable to this situation (see also Duncan (1957) and Kramer (1957) for this and additional extensions). The same general procedure as that for equal *n*'s is followed with only minor exceptions. To illustrate, consider some data selected from those reported by McGuigan, Crandell, and Suiter (1966). Subjects were asked to engage in one of five kinds of activities: Group 1 listened to a story presented by means of a tape recorder; Group 2 similarly listened to classical music; Group 3 memorized a portion of the same story that was visually presented to them; Group 4 read a selection from the story, and Group 5 listened to a blank tape on the tape recorder with instructions to pay attention in case they heard anything (which, of course, they didn't). Prior to these activities, all subjects rested

for one minute, following which they engaged in their activity for five minutes. Throughout the session several measures of covert behavior were recorded. The values that we shall here study are integrated electromyograms from the chin. The empirical question was whether or not covert oral behavior (chin electromyogram) was different under these various stimulating conditions. In particular, it was hypothesized that there would be greater chin activity during the presentation of language stimuli (Groups 1, 3, and 4) than under the nonlanguage stimulus (control) conditions (Groups 2 and 5). To approach the answer, the amount of chin activity during the rest condition was subtracted from the amount during each of the five minutes of the activity periods. The resulting curves for each of the groups is presented in Figure 9.11. There we can see that Group 1, who listened to the story, gave

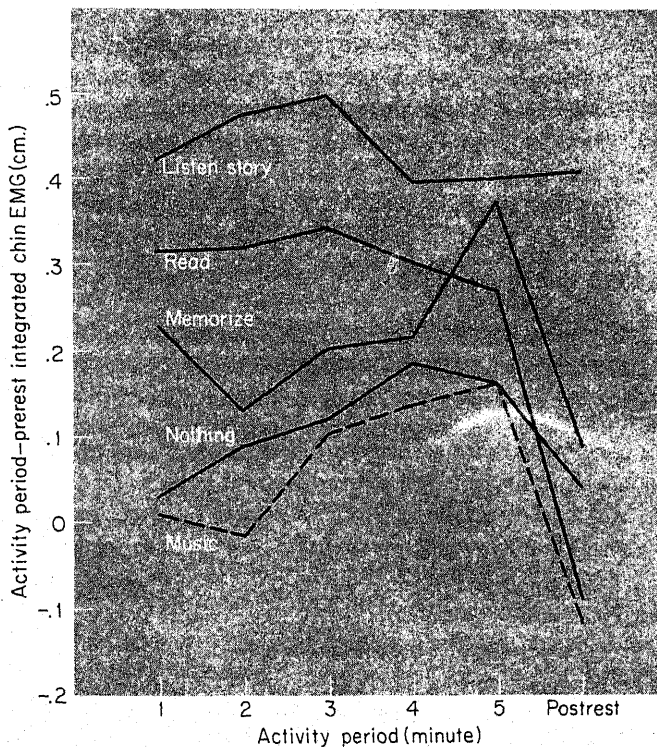


FIGURE 9.11.

Amplitude of covert oral behavior during five minutes of several kinds of activity. Mean chin EMG during rest was subtracted from values during each of the five minutes and plotted on the vertical axis.

the largest chin response throughout the five minute activity period. The next highest amounts of covert activity were given by Group 4 (reading), followed by Group 3, who memorized the story. The lowest levels of respond-

ing were made by Groups 5 and 2, the control groups. It is especially interesting to note that the subjects who received language stimuli (regardless of whether they listened, read, or memorized) yielded larger amounts of covert oral activity than did the subjects who did not receive language stimuli (the "nothing" and the "music" groups). Our question is, however, are these differences significant? The mean amount of covert oral behavior for each subject is presented in Table 9.7 [these values are the mean chin ampli-

Table 9.7. *Measures of Covert Oral Behavior During Various Activities.*

GROUP				
1 (Listen Story)	2 (Listen Music)	3 (Memorize)	4 (Read)	5 (Nothing)
1.7	-1.2	12.4	7.1	-2.2
-.6	2.0	2.4	-2.3	.6
28.9	.1	-.8	4.3	-.1
1.1	.5	0.0	0.0	.6
0.0	-.3	0.0	.6	3.1
-.5	3.7	2.2	2.4	2.1
2.1		.2	6.1	4.9
			8.1	.2
$\Sigma X = 32.70$	$\Sigma X = 4.80$	$\Sigma X = 16.40$	$\Sigma X = 26.30$	$\Sigma X = 9.20$
$\Sigma X^2 = 844.33$	$\Sigma X^2 = 19.48$	$\Sigma X^2 = 165.04$	$\Sigma X^2 = 183.13$	$\Sigma X^2 = 43.64$
$n = 7$	$n = 6$	$n = 7$	$n = 8$	$n = 8$
$\bar{X} = 4.67$	$\bar{X} = .80$	$\bar{X} = 2.34$	$\bar{X} = 3.29$	$\bar{X} = 1.15$

tude of each subject during the five minutes of activity subtracted from the mean chin amplitude during the resting (preactivity) period]. You can note that the numbers of subjects in the groups are different.

The first step is to compute ΣX , ΣX^2 , n , and \bar{X} for each group (see Table 9.7). With these values we shall compute the sum of squares of the five groups:

$$SS_1 = 844.33 - \frac{(32.70)^2}{7} = 691.58$$

$$SS_2 = 19.48 - \frac{(4.80)^2}{6} = 15.64$$

$$SS_3 = 165.04 - \frac{(16.40)^2}{7} = 126.62$$

$$SS_4 = 183.13 - \frac{(26.30)^2}{8} = 96.67$$

$$SS_5 = 43.64 - \frac{(9.20)^2}{8} = 33.06$$

We need next to compute s_e . However since the n 's of the groups are not equal, we shall have to use Equation (9.5) which, for five groups, becomes:

$$(9.7) \quad s_e = \sqrt{\frac{SS_1 + SS_2 + SS_3 + SS_4 + SS_5}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1) + (n_4 - 1) + (n_5 - 1)}}$$

Substituting the appropriate values of SS and n in Equation (9.7):

$$s_e = \sqrt{\frac{691.58 + 15.64 + 126.62 + 96.67 + 33.06}{(7 - 1) + (6 - 1) + (7 - 1) + (8 - 1) + (8 - 1)}} = 5.57$$

The df are computed as before:

$$df = 36 - 5 = 31$$

We now write down the values of r_p for five, four, three, and two groups. Assuming a 5 per cent level test we enter the column marked 5 in Table 9.2. Reading down the column we find the r_p for five groups that corresponds to 31 df . It is 3.20. For four groups $r_p = 3.12$ and so forth (Table 9.8).

Table 9.8. Values of r_p for a Five-Groups Design with 31 df .

	Groups			
	2	3	4	5
r_p	2.89	3.04	3.12	3.20

Up to this point the procedure for unequal n 's has been essentially the same as for equal n 's. We shall now depart to some extent. Instead of using Equation (9.4) we now use Equation (9.8) for computing R_p when we have unequal n 's:

$$(9.8) \quad R_p = (s_e)(r_p) \sqrt{\frac{1}{2} \left(\frac{1}{n_a} + \frac{1}{n_b} \right)}$$

Where n_a and n_b are the n 's for whatever two groups are being compared, as we shall shortly see. Let us start by ordering the means and testing the difference between the extreme groups:

	Group				
	2	5	3	4	1
Mean:	.80	1.15	2.34	3.29	4.67

The difference in means between Group 1 and Group 2 is 3.87. We next determine the value of R_p for this test. s_e will be the same for all tests, but r_p will depend on the number of groups being compared. In this comparison we are considering five means, so r_p will be that for comparing the extreme

means of five groups. From Table 9.8 we read: $r_s = 3.20$. The n 's are those for groups 1 and 2, i.e., 7 and 6 respectively. Hence, R_p is:

$$R_p = (5.57)(3.20)\sqrt{\frac{1}{2}\left(\frac{1}{7} + \frac{1}{6}\right)} = 17.82\sqrt{.1548} = 7.04$$

Since the mean difference between Groups 1 and 2 (3.87) does not exceed the R_p for this comparison (7.03), we may conclude that Group 1 is not significantly different from Group 2. We immediately know that, because the groups with the largest mean difference do not differ significantly, none of the other group means differ significantly; we need not conduct further tests. The statistical findings may be summarized as follows: The common line under all five groups indicates that there is no significant difference between any pair of them.

Group				
2	5	3	4	5
.80	1.15	2.34	3.29	4.67

Our conclusion from this study is that it has not been demonstrated that the various stimulus conditions produce significantly different amounts of covert oral behavior. The tendency, though, for the language conditions to produce larger amounts of covert oral behavior than the nonlanguage conditions is enticing and we should note that these findings were based on a relatively small number of subjects; this pilot investigation may be worth repeating with a larger n .

The above procedure can be extended to any number of groups. Besides substituting the appropriate n 's in Equation (9.8), one merely needs to be cautious about selecting the appropriate value of r_p .

ANALYSIS OF VARIANCE

Now having discussed Duncan's Range Test, we shall consider a second type of statistical procedure applied to the multi-groups design considered in this chapter. However, we shall attempt to explain only enough features of this procedure to allow you to apply it to your experimental work.¹²

As with the preceding designs, we set up a null hypothesis. Our statistical test will allow us to either reject the null hypothesis or fail to reject it. For the present design, consider the null hypothesis to be that there is no difference among the means of the several groups. Keeping this null hypothesis

¹²For a simplified but more thorough treatment of analysis of variance by psychologists you should read Spence, et al. (1968), or for more detailed treatments see the references cited on p. 206.

in mind, let us return to it after our consideration of analysis of variance.

You already have some acquaintance with the term variance (p. 181) which will help in the ensuing discussion. Review it now.

The simplest application of analysis of variance would be in testing the mean difference between two randomized groups. We have already discussed the *t*-test for this purpose. However, equivalent results would be obtained by conducting an analysis of variance on a two-groups design. That is, we could analyze a two-groups design by using either the *t*-test or the technique of analysis of variance and obtain precisely the same conclusions. The same statement cannot be made if more than two groups are used, for the obvious reason that the *t*-test cannot be used for testing more than two means simultaneously. Let us say that the dependent variable scores that result from a two-groups design are those plotted in Figure 9.12. That is,

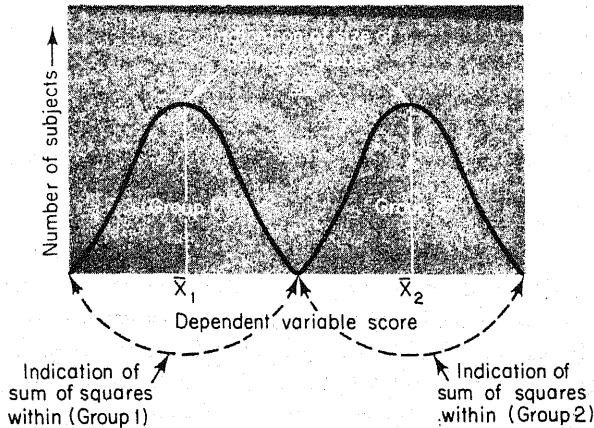


FIGURE 9.12.

A crude indication of the nature of within- and between-groups sum of squares using only two groups.

the curve to the left represents the scores made by the subjects in Group 1, and the frequency distribution to the right is for Group 2.

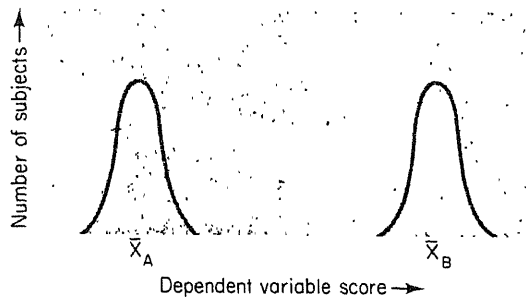
Now, are these groups significantly different? To answer this question by using analysis of variance we first need to note that the *total sum of squares* may be determined. The total sum of squares is a value that results when we take all subjects in the experiment into account as a whole. The total sum of squares is computed from the dependent variable scores of all the subjects, ignoring the fact that some were under one experimental condition while others were under another experimental condition. The important point for us to observe here is that the total sum of squares may be partitioned (analyzed) into parts. In particular, the total sum of squares may be parti-

tioned into two major components: the sum of squares *between groups* and the sum of squares *within groups*. Roughly, the sum of squares between groups may be thought of as determined by the extent to which the means of the two groups differ.

In Figure 9.12 the size of the between-groups sum of squares is crudely indicated by the distance between the two means. More accurately we may say that the larger the difference between the means, the larger is the between-groups sum of squares. The within-groups sum of squares, on the other hand, is determined by the extent to which the subjects in each group differ. If the subjects in Group 1 differ sizeably among themselves, and/or if the same is true for subjects in Group 2, the within-groups sum of squares is going to be large. And the larger the within-groups sum of squares, the larger the "error" in the experiment. By way of illustration, assume that all of the subjects in Group 1 have been treated precisely alike. Hence, if they were precisely alike when they went into the experiment, they should all receive the same score on the dependent variable. If this happened, the within-groups sum of squares (as far as Group 1 is concerned) would be zero, for there would be no variation among their scores. Of course, the within-groups sum of squares is almost never zero, since all the subjects are not the same before the experiment and the experimenter is never able to treat all of them precisely alike.

Let us now reason by analogy to the *t*-test. You will recall that the numerator of Equation (5.2) (p. 101) is a measure of the difference between the means of two groups. It is thus analogous to our between-groups sum of squares. The denominator of Equation (5.2) is a measure of the "error" in the experiment and is thus analogous to our within-groups sum of squares. This should be apparent when one notes that the denominator of Equation (5.2) is large if the variances of the groups are large, and small if the variances of the groups are small (see p. 186). Recall that the larger the numerator and the smaller the denominator of the *t* ratio, the greater the likelihood that the two groups are significantly different. The same is true in our analogy: the larger the between-groups sum of squares and the smaller the within-groups sum of squares, the more likely our groups are to be significantly different. Looking at Figure 9.12 we may say that the larger the distance between the two means and the smaller the within (internal) variances of the two groups, the more likely they are to be significantly different. For example, the difference between the means of the two groups of Figure 9.13 is more likely to be significant than the difference between the means of the two groups of Figure 9.12. This is so because the difference between the means in Figure 9.13 is represented as greater than that for Figure 9.12 and also because the sum of squares within the groups of Figure 9.13 is represented as less than for Figure 9.12.

We have discussed the case of two groups. Precisely the same general reasoning applies when there are more than two groups: the total sum of squares

**FIGURE 9.13.**

A more extreme difference between two groups than that shown in Figure 9.12. Here the between-groups sum of squares is greater but the within-groups sum of squares is less.

in the experiment is analyzed into two parts, the within- and the among-groups sum of squares. ("Between" is used for two groups, "among" is the same concept applied to more than two). If the difference among the several means is large, the among-groups sum of squares will be large. If the difference among the several means is small, the among-groups sum of squares will be small. If the subjects who are treated alike differ sizeably, then the within (internal) sum of squares of each group will be large. And if the individual group variances are large, the within-group sum of squares will be large. The larger the among-groups sum of squares and the smaller the within-group sum of squares the more likely it is that the groups differ significantly.

We have attempted to present, in a surface fashion, the major rationale underlying analysis of variance. As we now turn to the computation of the several sums of squares we shall be more precise. The equations to be given are based on the following reasoning and their computation automatically accomplishes what we are going to say. First, a mean is computed that is based on all of the dependent variable values in the experiment taken together (ignoring the fact that some subjects were under one condition and others under another condition). Then, the total sums of squares measures the deviation of all of the scores from this overall mean. The among groups sum of squares is a measure of the deviation of the means of the several groups from the overall mean. And the within groups sum of squares is a pooled sum of squares based on the deviation of the scores in each group from the mean of that group. As we proceed we shall continue to enlarge on these introductory statements.

Our purpose will be to compute the total *SS* and then analyze it into its parts. A generalized equation for computing the total *SS* is:

$$(9.9) \quad \text{Total } SS = (\sum X_1^2 + \sum X_2^2 + \cdots + \sum X_r^2) - \frac{(\sum X_1 + \sum X_2 + \sum X_3 + \cdots + \sum X_r)^2}{N}$$

As before, the subscript r simply indicates that we continue adding the values indicated (the sum of X -squares, and the sum of X respectively) for as many groups as we have in the experiment.

Our next step is to analyze the total SS into components. There are two major components, that among groups and that within groups. A generalized equation for computing the among-groups SS is:

$$(9.10) \quad \text{Among } SS = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \cdots + \frac{(\Sigma X_r)^2}{n_r} - \frac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \cdots + \Sigma X_r)^2}{N}$$

The within-groups component of the total SS may be computed by subtraction. That is:

$$(9.11) \quad \text{Within } SS = \text{Total } SS - \text{Among } SS$$

In a more-than-two-randomized-groups design, of course, there may be any number of groups. To compute the several SS we must compute the ΣX and ΣX^2 separately for each group. The subscripts, as before, indicate the different groups. Hence ΣX_1 is the sum of the dependent variable scores for Group 1, ΣX_3^2 is the sum of the squares of the dependent variable scores for Group 3, and so forth. N remains the total number of subjects in the experiment. To illustrate the analysis of variance procedure, consider an experiment related to one previously analyzed (pp. 101-110). In this study Jacobson, Fried, and Horowitz (1966) classically conditioned one group of planarians to a light; more specifically this group (Group CC for "classically conditioned") received paired presentations of a light and a shock. They normally contract when shocked, but after conditioning they also contracted to the conditioned stimulus, the light. Group PC (for "pseudoconditioning") was treated in the same way as Group CC, but the light and shock were not paired, i.e., these planarians were shocked and received light on their trials, but the light and shock were not associated so that one would not expect conditioning to occur. The third group (NC for "nonconditioned"), simply rested in their home containers and were not exposed to the experimental situation.

After the above procedure was followed, untrained planarians were injected with ribonucleic acid (RNA, cf., p. 101) from the above groups. More specifically, a new group of planarians received RNA that was extracted from Group CC, a second naive group received injections of RNA from Group PC, and a third with injections from Group NC. These new groups were then tested to see how often they would give the conditioned response (contraction) to the conditioned stimulus (light). The number of conditioned responses made by each animal during 25 test trials is presented in Table 9.9.

To compute the total SS , we may write the specialized form of Equation (9.9) for three groups as follows:

$$(9.12) \quad \text{Total } SS = (\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2) - \frac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3)^2}{N}$$

We can see that the sum of X for Group NC is 40, for Group PC it is 46, and for Group CC it is 205. Or written in terms of Equation (9.12) we may say that $\Sigma X_1 = 40$, $\Sigma X_2 = 46$, and $\Sigma X_3 = 205$. Similarly, $\Sigma X_1^2 = 104$, $\Sigma X_2^2 = 154$, $\Sigma X_3^2 = 1721$, and $N = 75$. Substituting these values in Equation (9.12), we find the total SS to be:

$$\text{Total } SS = (104 + 154 + 1721) - \frac{(40 + 46 + 205)^2}{75} = 849.92$$

Table 9.9. *Number of Responses on the 25 Test Trials for each Injected Planarian.*

PLANARIANS INJECTED WITH RNA FROM:			
Group 1 (NC)	Group 2 (PC)	Group 3 (CC)	
0	0	6	
0	0	6	
0	0	6	
0	0	7	
1	0	7	
1	0	7	
1	0	7	
1	0	7	
1	0	8	
1	1	8	
1	1	8	
1	2	8	
1	2	8	
1	2	9	
1	2	9	
2	3	9	
2	3	9	
2	3	9	
2	3	9	
3	3	9	
3	3	9	
3	4	10	
3	4	10	
4	5	10	
5	5	10	
$\Sigma X:$	40	46	205
$\Sigma X^2:$	104	154	1721
$n:$	25	25	25
$\bar{X}:$	1.60	1.84	8.20

To compute the among-groups SS for three groups, we substitute the appropriate values in Equation (9.13), the specialized form of Equation

(9.10) for three groups. This requires that we merely substitute the value of ΣX for each group, square it, and divide by the number of subjects in each group. The last term, we may note, is the same as the last term in Equation (9.12). For this reason it is not necessary to compute it again, providing there was no error in its computation the first time. Making the appropriate substitutions from Table 9.9, and performing the indicated computations we find that:

$$\begin{aligned}
 (9.13) \quad \text{Among-groups } SS &= \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} - \frac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3)^2}{N} \\
 &= \frac{(40)^2}{25} + \frac{(46)^2}{25} + \frac{(205)^2}{25} - 1129.08 = 700.56
 \end{aligned}$$

Substituting in Equation (9.11), we find that:

$$\text{Within-groups } SS = 849.92 - 700.56 = 149.36$$

We have said that we are conducting an analysis of variance and you may have wondered where the variances are that we are analyzing. We shall consider them now, but under a different name, for they are referred to in our sample values not as variances, but as *mean squares*. That is, we are computing sample values as estimates of population values. The mean squares (sample values) are estimates of the variances (population values). For example, the mean square within groups is an estimate of the within groups variance. The rule for computing mean squares is simple: divide a given sum of squares by the appropriate degrees of freedom.

In introducing the equations that we use to determine the three degrees of freedom that we need, let us emphasize what we have done with regard to sums of squares. We have computed a total SS and partitioned it into two parts, the among SS and the within SS . The same procedure is followed for df . First we determine that:

$$(9.14) \quad \text{Total } df = N - 1$$

then that:

$$(9.15) \quad \text{Among } df = r - 1$$

and that:

$$(9.16) \quad \text{Within } df = N - r$$

For our example we then find that, with $N = 75$, and $r = 3$,

$$\text{Total } df = 75 - 1 = 74$$

$$\text{Among } df = 3 - 1 = 2$$

$$\text{Within } df = 75 - 3 = 72$$

And we may note that the among df plus the within df equals the total df ($72 + 2 = 74$).

There are two mean squares that we need to compute, a mean square for the among-groups source of variation, and that for the within-groups. To compute the former we divide the among-groups SS by the among-groups df , and similarly for the latter. Hence, the within-groups mean square is 149.36 divided by 72. We shall enter these values in a summary table (Table 9.10).

Table 9.10 *Summary Table for an Analysis of Variance.*

<i>Source of Variation</i>	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>
Among Groups	700.56	2	350.28	169
Within Groups	149.36	72	2.07	
Total	849.92	74		

Now, as we have previously indicated, if the among-groups mean square is sizeable, relative to the within-groups mean square, then we may conclude that the dependent variable values for the groups are different.

However, we must again face the problem: how sizeable is "sizeable," i.e., how large must the among component be in order for us to conclude that a given independent variable is effective? To answer this we apply a suitable statistical test. The test that is considered most appropriate is the F -test, which was developed by Professor Sir Ronald Aymer Fisher, one of the outstanding statisticians of all time. It was so named in his honor by another outstanding statistician, Professor George W. Snedecor. The F statistic for this design may be defined as follows:

$$(9.17) \quad F = \frac{\text{Mean square among groups}}{\text{Mean square within groups}}$$

This statistic is obviously easy to compute, and we may note the similarity between it and the t -test. For in both cases the numerator is an indication of the differences between or among groups (plus experimental error) and the denominator is an indication of the experimental error, or as it is also called the error variance of the experiment. More particularly, in the simplest applications of the F -test, the numerator contains an estimate of the error variance plus an estimate of the "real" effect (if any) of the independent variable. The denominator is only an estimate of the error variance. Now you can see what happens when you divide the numerator by the denominator: The computed value of F reflects the effect that the independent variable had in producing a difference between means. For example, suppose the independent variable is totally ineffective in influencing the dependent variable. In this case we would expect (at least in the long run) that the numerator would not contain any contribution from the independent vari-

able (there would be no "real" among-groups mean square). Hence, the value for the numerator would only be an estimate of the error variance; a similar estimate of the error variance is in the denominator. Therefore, in the long run (with a large number of df), if you divide a value for the error variance by a value for the error variance, you obtain an F in the neighborhood of 1.0.

Thus, any time we obtain an F of approximately one, we can be rather sure that variation of the independent variable did not produce a difference in the dependent variable means of our groups. However, the numerator may be somewhat larger than the denominator — the among-groups mean square may be somewhat larger than the within-groups mean square. Now the question is, how large must the numerator of the F ratio be, before we conclude that the means of the groups are really different. For, if the numerator is large (relative to the denominator), the value of F will be large. You will note that this is the same question we asked concerning the t -test. That is, how large must t be before we can reject our null hypothesis? And we shall answer this question for the F -test in a manner similar to that for the t -test. First, however, we should actually compute our values of F . Following our example, we have divided the mean square within groups (2.07) into the mean square between groups (350.28), and inserted the resulting value (169) of F in Table 9.10.

Just as with the t -test, we next must determine the value of P that is associated with our computed value of F . Assuming that we have set a significance level of 0.05, if the value of F has a probability of less than 0.05, then we may reject the null hypothesis; we may assert that there is a significant difference among the means of the groups. If, however, the P associated with our F is larger than 0.05, then we fail to reject the null hypothesis. We conclude that there is no significant difference among the group means.

To ascertain the value of P associated with our F , refer to Table 9.11, which fulfills the same function as the Table of t , although it is a bit different to use. Let us initially note that: (1) across the top we find " df associated with the numerator" and (2) down the left side we find " df associated with the denominator." Therefore, we know that we need two df values to enter the table of F . In this example we have 2 df for among groups (the numerator of the F -test) and 72 df for within groups (the denominator of the F -test). Hence we find the column labeled "2" and read down to find a row labeled "72." There is none, but there are rows for 60 and 120 df ; 72 falls between these two values. We find a row for a P of 0.01, a row for a P of 0.05, and rows for P 's of 0.10 and 0.20. We are making a 0.05 level test, so we shall ignore the other values of P . With 2 and 72 df , we interpolate between 3.15 and 3.07 and find that we must have an F of 3.13 for significance at the 5 per cent level. Since the computed F (169.22) exceeds this value,¹³ the null

¹³A bit of an understatement.

hypothesis is rejected — we conclude that the groups differ significantly. And, on the assumption that proper experimental techniques have obtained, it is concluded that variation of the independent variable significantly influenced the dependent variable. More specifically, with regard to this experiment, injection of RNA from groups with various training resulted in the groups differing significantly on the dependent variable measure.

Just as with the table of t , you should study the table of F sufficiently to make sure that you have an adequate understanding of it. To provide a little practice, say that you have six groups in your experiment with ten subjects per group. In this event you have five degrees of freedom for your among source of variation and 54 df for the within. Assume a 1 per cent level test. What is the value of F that you must obtain in order to reject the null hypothesis? To answer this question, enter the column labeled “5” and read down until you find the rows for 54 df . There is no row for 54 df so you must interpolate. The 54 df falls between the tabled values of 40 df and 60 df . If you had had 40 df for your within groups, then you would have needed a computed F of 3.51 in order to reject the null hypothesis; similarly if you had had 60 df , you would have required an F of 3.34. By linearly interpolating we find that an F of 3.39 is required for significance at the 1 per cent level. Try some additional problems for yourself.

Now, if you had conducted the above experiment you might feel quite happy with yourself; you would have succeeded in rejecting the null hypothesis. But wait a moment. What null hypothesis did you reject? Your conclusion is that there is a (at least one) difference among your groups. But where does the difference lie? Is it between Groups 1 and 2, between Groups 1 and 3, between Groups 2 and 3, or are two, or all, of these differences significant? The conventional (but, as we shall see, inappropriate) answer to this question is to run t -tests between the groups as indicated above; this involves running three t -tests. Using the data in Table 9.9, compute the values of t ; you will find them to be:

Between Groups 1 and 2: t_{12}	= .58
Between Groups 1 and 3: t_{13}	= 18.85
Between Groups 2 and 3: t_{23}	= 18.17

Assuming a significance level of 0.05, we find that the t between Groups 1 and 3 and between Groups 2 and 3 is significant; our question as to where among the three groups the significant difference lies is answered.¹⁴ That is, Group 3 (the Group that received RNA from the planarians that were classically conditioned) yielded significantly more contractions to the light during the test trials than did the other two (control) groups. The “. . . data would seem to suggest that a specific learned response was transferred by way of the injection of the RNA preparation” (Jacobson et al., 1966, p. 5).

¹⁴Obviously we have now tested three additional null hypotheses set up for our three t -tests, e.g., there is no difference between the means of Groups 1 and 2.

Table 9.11. Table of F .

df Associated with Denominator	df ASSOCIATED WITH NUMERATOR										
	P	1	2	3	4	5	6	8	72	24	∞
1	0.01	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
	0.05	161.45	199.50	215.71	224.58	230.16	233.99	238.88	243.91	249.05	254.32
	0.10	39.86	49.50	53.59	55.83	57.24	58.20	59.44	60.70	62.00	63.33
	0.20	9.47	12.00	13.06	13.73	14.01	14.26	14.59	14.90	15.24	15.58
2	0.01	98.49	99.00	99.17	99.25	99.30	99.33	99.36	99.42	99.46	99.50
	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
	0.10	8.53	9.00	9.16	9.24	9.29	9.33	9.37	9.41	9.45	9.49
	0.20	3.56	4.00	4.16	4.24	4.28	4.32	4.36	4.40	4.44	4.48
3	0.01	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
	0.10	5.54	5.46	5.39	5.34	5.31	5.28	5.25	5.22	5.18	5.13
	0.20	2.68	2.89	2.94	2.96	2.97	2.97	2.98	2.98	2.98	2.98
4	0.01	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
	0.10	4.54	4.32	4.19	4.11	4.05	4.01	3.95	3.90	3.83	3.76
	0.20	2.35	2.47	2.48	2.48	2.48	2.47	2.47	2.46	2.44	2.43
5	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
	0.10	4.06	3.78	3.62	3.52	3.45	3.40	3.34	3.27	3.19	3.10
	0.20	2.18	2.26	2.25	2.24	2.23	2.22	2.20	2.18	2.16	2.13
6	0.01	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
	0.10	3.78	3.46	3.29	3.18	3.11	3.05	2.98	2.90	2.82	2.72
	0.20	2.07	2.13	2.11	2.09	2.08	2.06	2.04	2.02	1.99	1.95

7	0.01	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
	0.10	3.59	3.26	3.07	2.96	2.88	2.83	2.75	2.67	2.58	2.47
	0.20	2.00	2.04	2.02	1.99	1.97	1.96	1.93	1.91	1.87	1.83
8	0.01	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
	0.10	3.46	3.11	2.92	2.81	2.73	2.67	2.59	2.50	2.40	2.29
	0.20	1.95	1.98	1.95	1.92	1.90	1.88	1.86	1.83	1.79	1.74
9	0.01	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
	0.10	3.36	3.01	2.81	2.69	2.61	2.55	2.47	2.38	2.28	2.16
	0.20	1.91	1.94	1.90	1.87	1.85	1.83	1.80	1.76	1.72	1.67
10	0.01	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
	0.10	3.28	2.92	2.73	2.61	2.52	2.46	2.38	2.28	2.18	2.06
	0.20	1.88	1.90	1.86	1.83	1.80	1.78	1.75	1.72	1.67	1.62
11	0.01	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
	0.05	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
	0.10	3.23	2.86	2.66	2.54	2.45	2.39	2.30	2.21	2.10	1.97
	0.20	1.86	1.87	1.83	1.80	1.77	1.75	1.72	1.68	1.63	1.57
12	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
	0.05	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
	0.10	3.18	2.81	2.61	2.48	2.39	2.33	2.24	2.15	2.04	1.90
	0.20	1.84	1.85	1.80	1.77	1.74	1.72	1.69	1.65	1.60	1.54
13	0.01	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
	0.05	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
	0.10	3.14	2.76	2.56	2.43	2.35	2.28	2.20	2.10	1.98	1.85
	0.20	1.82	1.88	1.78	1.75	1.72	1.69	1.66	1.62	1.57	1.51

Tabl 9.11.* (Continued)

df Associated with Denominator	df ASSOCIATED WITH NUMERATOR										
	P	1	2	3	4	5	6	8	12	24	∞
14	0.01	8.86	6.51	5.56	5.08	4.69	4.46	4.14	3.80	3.43	3.00
	0.05	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
	0.10	3.10	2.73	2.52	2.39	2.31	2.24	2.15	2.05	1.94	1.80
	0.20	1.81	1.81	1.76	1.78	1.70	1.67	1.64	1.60	1.55	1.48
15	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
	0.10	3.07	2.70	2.49	2.36	2.27	2.21	2.12	2.02	1.90	1.76
	0.20	1.80	1.79	1.75	1.71	1.68	1.66	1.62	1.58	1.53	1.46
16	0.01	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
	0.05	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
	0.10	3.05	2.67	2.46	2.33	2.24	2.18	2.09	1.99	1.87	1.72
	0.20	1.79	1.78	1.74	1.70	1.67	1.64	1.61	1.56	1.51	1.43
17	0.01	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
	0.05	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
	0.10	3.03	2.64	2.44	2.31	2.22	2.15	2.06	1.96	1.84	1.69
	0.20	1.78	1.77	1.72	1.68	1.65	1.63	1.59	1.55	1.49	1.42
18	0.01	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
	0.05	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
	0.10	3.01	2.62	2.42	2.29	2.20	2.13	2.04	1.93	1.81	1.66
	0.20	1.77	1.76	1.71	1.67	1.64	1.62	1.58	1.53	1.48	1.40
19	0.01	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
	0.05	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
	0.10	2.99	2.61	2.40	2.27	2.18	2.11	2.02	1.91	1.79	1.63
	0.20	1.76	1.75	1.70	1.66	1.63	1.61	1.57	1.52	1.46	1.39

20	0.01	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
	0.10	2.97	2.59	2.38	2.25	2.16	2.09	2.00	1.89	1.77	1.61
	0.20	1.76	1.75	1.70	1.65	1.62	1.60	1.56	1.51	1.45	1.37
21	0.01	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
	0.05	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
	0.10	2.96	2.57	2.36	2.23	2.14	2.08	1.98	1.88	1.75	1.59
	0.20	1.75	1.74	1.69	1.65	1.61	1.59	1.55	1.50	1.44	1.36
22	0.01	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
	0.05	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
	0.10	2.95	2.56	2.35	2.22	2.13	2.06	1.97	1.86	1.73	1.57
	0.20	1.75	1.73	1.68	1.64	1.61	1.58	1.54	1.49	1.43	1.35
23	0.01	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
	0.05	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
	0.10	2.94	2.55	2.34	2.21	2.11	2.05	1.95	1.84	1.72	1.55
	0.20	1.74	1.73	1.68	1.63	1.60	1.57	1.53	1.49	1.42	1.34
24	0.01	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
	0.05	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
	0.10	2.93	2.54	2.33	2.19	2.10	2.04	1.94	1.83	1.70	1.53
	0.20	1.74	1.72	1.67	1.63	1.59	1.57	1.53	1.48	1.42	1.33
25	0.01	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
	0.05	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
	0.10	2.92	2.53	2.32	2.18	2.09	2.02	1.93	1.82	1.69	1.52
	0.20	1.73	1.72	1.66	1.62	1.59	1.56	1.52	1.47	1.41	1.32
26	0.01	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
	0.05	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
	0.10	2.91	2.52	2.31	2.17	2.08	2.01	1.92	1.81	1.68	1.50
	0.20	1.73	1.71	1.66	1.62	1.58	1.56	1.52	1.47	1.40	1.31

Table 9.11* (Continued)

<i>df</i> Associated with Denominator	<i>P</i>	1	2	3	4	5	6	8	12	24	∞
27	0.01	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
	0.05	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
	0.10	2.90	2.51	2.30	2.17	2.07	2.00	1.91	1.80	1.67	1.49
	0.20	1.73	1.71	1.66	1.61	1.58	1.55	1.51	1.46	1.40	1.30
28	0.01	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
	0.05	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
	0.10	2.89	2.50	2.29	2.16	2.06	2.00	1.90	1.79	1.66	1.48
	0.20	1.72	1.71	1.65	1.61	1.57	1.55	1.51	1.46	1.39	1.30
29	0.01	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
	0.05	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
	0.10	2.89	2.50	2.28	2.15	2.06	1.99	1.89	1.78	1.65	1.47
	0.20	1.72	1.70	1.65	1.60	1.57	1.54	1.50	1.45	1.39	1.29
30	0.01	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
	0.10	2.88	2.49	2.28	2.14	2.05	1.98	1.88	1.77	1.64	1.46
	0.20	1.72	1.70	1.64	1.60	1.57	1.54	1.50	1.45	1.38	1.28
40	0.01	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
	0.05	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
	0.10	2.84	2.44	2.23	2.09	2.00	1.93	1.83	1.71	1.57	1.38
	0.20	1.70	1.68	1.62	1.57	1.54	1.51	1.47	1.41	1.34	1.24
60	0.01	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
	0.05	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
	0.10	2.79	2.39	2.18	2.04	1.95	1.87	1.77	1.66	1.51	1.29
	0.20	1.68	1.65	1.59	1.55	1.51	1.48	1.44	1.38	1.31	1.18

To briefly summarize the general approach of this section, we may say that on the assumption that only comparisons between pairs are to be made, and on the assumption that all possible pairs are to be compared, we first run an F -test. If the value of this "over-all" F is significant, we then conduct all possible t -tests in order to ascertain which specific groups are significantly different. If, however, the over-all F is not significant, then we conclude that there are no significant differences between the various pairs of groups, thereby not running additional t -tests.

Let us briefly comment on a different approach that you might wish to take in an experiment. Suppose that you are not really interested in making all possible comparisons between your groups. For instance, suppose that you conduct a three-group experiment, and that your empirical hypothesis suggests that Groups 1 and 2 should differ and that Groups 1 and 3 should also differ; but in this event the comparison between Groups 2 and 3 is rather uninteresting to you; your hypothesis says nothing about this comparison. In this event you need not conduct your over-all F -test; you go directly to your t -test analysis, computing the two indicated values of t .

APPENDIX TO CHAPTER 9

To illustrate the inappropriateness of applying the t -test to the multi-randomized groups design, suppose that we run a two-groups experiment on rats. We set our significance level at 0.05. Recall that, assuming that the null hypothesis is true, this significance level means that if we obtain a t that has a P of 0.05, the odds are five in 100 that a t of this size or larger could have occurred by chance. Since this would happen only rarely (5 per cent of the time) we reason that the t was not the result of random fluctuations. Rather, we prefer to conclude that the two groups are "really" different as measured by the dependent variable. We thus reject our null hypothesis and conclude that variation of the independent variable was effective in producing the difference between our two groups. Now, after completing the above work, say that we conduct a new two-groups experiment, for example one on schizophrenics. Note that the two experiments are independent of each other. In the experiment on schizophrenics we also set our significance level at 0.05, and follow the same procedure as before. Again our significance level means that the odds are five in 100 that a t of the corresponding size could have occurred by chance.

But let us ask a question. Given a significance level of 0.05 in each of the two experiments, what are the odds that by chance the t in one, the other, or both experiments will be significant? Before you reach a hasty conclusion, let us caution you that the probability is *not* 0.05. Rather, the joint prob-

ability could be shown to be 0.0975.¹⁵ That is, the odds of obtaining a t significant at the 0.05 level in either or both experiments are 975 out of 10,000. And this is certainly different from 0.05.

To illustrate, we might develop an analogy: What is the probability of obtaining a head in two tosses of a coin? On the first toss it is one in two, and on the second toss it is one in two. But the probability of obtaining two heads on two successive tosses (before your first toss) is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. To develop the analogy further, the probability of obtaining a head on the first toss, or on the second toss, or on both tosses (again, computed before *any* tosses) is $P = 0.75$.

In a different situation, suppose that we conduct a three-groups experiment. In this case there are three t -tests in which we would probably be interested: a t between Groups 1 and 2, between Groups 1 and 3, and between Groups 2 and 3. Assume that we set a significance level of 0.05 *for each* t . If the first t -test yields a value significant beyond the 0.05 level, we reject the null hypothesis. And likewise for the other two t -tests. But what are the odds of obtaining a significant t when we consider all three t -tests? That is, what are the odds of obtaining a significant t in at least one of the following situations:

- First: Between Groups 1 and 2
- or Second: Between Groups 1 and 3
- or Third: Between Groups 2 and 3
- or Fourth: Between Groups 1 and 2 and also between Groups 1 and 3
- or Fifth: Between Groups 1 and 2 and also between Groups 2 and 3
- or Sixth: Between Groups 1 and 3 and also between Groups 2 and 3
- or Seventh: Between Groups 1 and 2 and also between Groups 2 and 3 and also between Groups 1 and 3.

The answer to this question is more complex than before. The reason for this is that these t -tests are not independent. For example, the computed value of t between Groups 1 and 2 would be related to the computed value of t between Groups 1 and 3 because Group 1 occurs in both t -tests. Similarly, t -tests between Groups 2 and 3 and between Groups 1 and 3 would not be independent. And lacking independence in the t -tests, it would be difficult to say just what the joint or over-all significance level is (as we were able to say in the previous case that it was 0.0975). About the best we can say for the general case is that the significance level for all possible t -tests is less than that which would obtain if the t -tests were independent. This is not much help. But the moral is: The running of all possible t -tests (between pairs of

¹⁵By the following formula: $P_j = 1 - (1 - a)^k$ where P_j is the joint probability, a is the significance level, and k is the number of independent experiments. For instance in this case $a = .05$, $k = 2$. Therefore $P_j = 1 - (1 - 0.05)^2 = 0.0975$.

means) is not a satisfactory technique of statistical analysis. It simply does not provide a reasonable level of significance when all possible t 's are considered. However, this is not the worst of it. If, for instance, we had seven groups in our experiment, we would have to run 21 t -tests in order to consider all possible combinations between pairs of means. Although we would not shirk the work involved in running all these tests if it were necessary, we still have a limited amount of energy and would prefer to expend it in ways other than running a great many t -tests, if possible.

One appropriate solution might be to run fewer t -tests. Assuming that we are only interested in t -tests between pairs of means, we could select those t -tests that are independent and run them. Following the seven-groups example, however, it could be shown that only three such t -tests *between pairs* would be independent. And we are usually interested in comparing more than three pairs of groups in such an experiment.

The importance of this discussion is that we have demonstrated our objections to the frequently used procedure of analyzing the multi-randomized-groups design, that of an analysis of variance followed by all possible t -tests. These criticisms are not directed toward the analysis of variance phase, for that by itself is perfectly legitimate. Thus you may conduct your analysis of variance and run your F -test. If it is significant, then you know that some significant difference exists among your groups, but that is all that the F -test tells you, for you do not know where the difference lies.

Duncan's Range Test seems considerably more appropriate, for it: (1) allows us to make all possible comparisons between pairs of our groups, just as the 21 t -tests in the previous example would, (2) is less work than running an F -test and a large number of t -tests; and (3) provides a more reasonable level of significance than for all possible t -tests, considered jointly.¹⁶

**SUMMARY OF THE COMPUTATION OF
DUNCAN'S RANGE TEST FOR A
RANDOMIZED-GROUPS DESIGN WITH
MORE THAN TWO GROUPS**

Assume that the following dependent variable scores have been obtained for four groups of subjects.

Group 1	Group 2	Group 3	Group 4
1	2	8	7
1	3	8	8
3	4	9	9
5	5	10	9
5	6	11	10
6	6	12	11
7	6	12	11

¹⁶We should point out, too, that there are other approaches than those that we have presented (cf. Ryan, 1959; Gaito, 1959).

1. First we wish to compute ΣX , ΣX^2 , n , and \bar{X} values for each group.

ΣX :	28	32	70	65
ΣX^2 :	146	162	718	617
n :	7	7	7	7
\bar{X} :	4.00	4.57	10.0	9.29

2. Compute the sum of squares (SS) for each group. These values are determined by substituting in Equation (9.1) and performing the indicated operations.

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

$$SS_1 = 146 - \frac{(28)^2}{7} = 34.00$$

$$SS_2 = 162 - \frac{(32)^2}{7} = 15.71$$

$$SS_3 = 718 - \frac{(70)^2}{7} = 18.00$$

$$SS_4 = 617 - \frac{(65)^2}{7} = 13.43$$

3. Using Equation (9.5), compute the square root of the error variance. Substituting the above values and performing the appropriate operations, we find that:

$$s_e = \sqrt{\frac{SS_1 + SS_2 + SS_3 + SS_4}{4(n-1)}} = 1.84$$

4. Determine the degrees of freedom for Duncan's Range Test, where:

$$df = N - r = 28 - 4 = 24$$

5. Assuming a 1 per cent level test, enter Table 9.6 to determine the appropriate values of r_p . Since we have four means in the present example, we need to enter Table 9.6 at the columns labeled 4, 3, and 2. With 24 df we find the values of r_p to be:

	Number of Groups		
	2	3	4
r_p	3.96	4.14	4.24

6. Compute the least significant ranges (R_p) for comparisons between two groups, among three groups, and among four groups. The equation [Equation (9.4)] is:

$$R_p = (s_e) (r_p) \sqrt{1/n}$$

Making the appropriate substitutions to determine R_p for two groups (i.e., R_2), and performing the indicated operations:

$$R_2 = (1.84) (3.96) \sqrt{1/7} = 2.75$$

Similar substitutions and performance of the operations result in the values of R_p for three and for four groups:

$$R_3 = (1.84) (4.14) \sqrt{1/7} = 2.88$$

$$R_4 = (1.84) (4.24) \sqrt{1/7} = 2.95$$

The computed values of R_p may now be summarized:

	<i>Number of Groups</i>		
	2	3	4
R_p :	2.76	2.90	2.96

7. The next step is to rank the means of the groups from lowest to highest.

	<i>Group 1</i>	<i>Group 2</i>	<i>Group 4</i>	<i>Group 3</i>
\bar{X}	4.00	4.57	9.29	10.00

8. We now test for significant differences among the various pairs of means. Starting with the highest (Group 3) and the lowest (Group 1), we can see that the difference between their means is 6.00. Determining the appropriate value of R_p for the comparison (that for four groups, hence $R_4 = 2.96$) we compare the mean difference with the value of R_p . In this case 6.00 exceeds 2.96. Therefore, the means of Groups 1 and 3 differ significantly. The next comparison is between the highest group (Group 3) and the second from the lowest group (Group 2). The difference between these pairs of means is 5.43. The value of R_p for comparing three groups is 2.90. Since 5.43 exceeds 2.90, the means of Groups 2 and 3 differ significantly. The next comparison is between the highest and the next to highest means. The mean difference is 0.71. This is a two-group comparison; hence $R_p = 2.76$. Since the difference between the means of Groups 3 and 4 (.71) does not exceed 2.76, these two groups do not differ significantly. We now test for a significant difference between the next to highest mean and the lowest mean. The mean difference is 5.29, and $R_3 = 2.90$. Therefore, Groups 4 and 1 differ significantly. In the test between Groups 4 and 2 the mean difference is 4.72 and $R_2 = 2.76$. Therefore Groups 4 and 2 differ significantly. The final comparison is between Groups 1 and 2. Their mean difference is .57. Since $R_2 = 2.76$, these two groups do not differ significantly.

9. We now summarize the results of our tests of significance. We found, in our example, that significant differences did not exist between Groups 1 and

2, nor between Groups 3 and 4. All other differences were significant. These findings are summarized as follows:

<i>Group 1</i>	<i>Group 2</i>	<i>Group 4</i>	<i>Group 3</i>
4.00	4.57	9.29	10.00

PROBLEMS

1. An experimenter was interested in assessing the relative sociability scores of different majors in his college. He selected a random sample of students who were majoring in English, Art, and Chemistry and administered a standardized test of sociability. Assuming a 1 per cent level of significance, did the three groups differ on this measure? Can it be said that all three groups are significantly different from each other?

SOCIABILITY SCORES

<i>English Majors</i>	<i>Art Majors</i>	<i>Chemistry Majors</i>
1,2,2,2,3,3	5,5,5,6,6,6	9,9,9,9,10,10

2. A Physical Education professor is interested in the effect of practice on the frequency of making goals in hockey. After consulting a psychologist he designs the following experiment. Four groups are formed such that Group 1 received the most practice, Group 2 the second most practice, Group 3 the third most practice, and Group 4 the least amount of practice. Dependent variable scores represent the number of goals made by each subject during a test period. Determine which groups are significantly different. (Significance level is 0.05.)

<i>Group</i>			
I	II	III	IV
5	1	5	10
7	4	0	9
9	0	2	6
7	0	1	5
6	8	1	8
5	3	4	9
9	2	3	8
2	1	0	2

3. An experimenter was interested in testing the hypothesis that the greater the hunger drive, the more correct choices a rat would make in a certain number of runs in a maze. He formed five groups of rats such that Group 1 had zero hours of food deprivation, Group 2 had 12 hours, Group 3

had 24 hours of food deprivation, Group 4 had 36 hours of food deprivation, and Group 5 had 48 hours of food deprivation. Setting a 5 per cent level of significance, was his hypothesis confirmed?

NUMBER OF CORRECT CHOICES				
<i>Group 1</i>	<i>Group 2</i>	<i>Group 3</i>	<i>Group 4</i>	<i>Group 5</i>
0	1	0	3	4
0	1	1	3	5
1	3	1	4	6
2	3	2	5	7
3	4	4	6	7
3	4	4	7	8
4	4	5	7	9
5	4	5	8	10
6	5	7	9	11
7	6	8	10	12
7	7	9	11	14

4. An experiment is conducted to determine which of three methods of teaching Spanish is superior. Assuming that the experiment has been adequately conducted, that the 5 per cent level of significance has been set, and that the higher the test score the better the performance after training on the three methods, which method is to be preferred?

<i>Method A Subjects</i>	<i>Method B Subjects</i>	<i>Method C Subjects</i>
55	46	45
52	40	41
50	35	37
48	32	36
47	31	30
46	28	25
40	25	24
35	22	21
	21	21
	19	20
		19
		18
		17

EXPERIMENTAL DESIGN

The Factorial Design

All of the preceding designs are appropriate to the investigation of a single independent variable. If the dependent variable is varied in two ways, one of the two-groups designs is used. If the independent variable is varied in more than two ways, the multi-group design is used. It is possible, however, to study more than one independent variable in a single experiment. One possible design for studying two or more independent variables in a single experiment is the *factorial design*. A complete factorial design is one where all possible combinations of the selected values of each of the independent variables are used. To illustrate a simple factorial design, let us consider an experiment by Harley and Harley (1966) on learning during hypnosis. Among the independent variables they studied were: (1) whether or not the subjects were hypnotized; and (2) susceptibility to be hypnotized. These are both continuous variables, but two values of each were selected for study so that the first was dichotomized as above, and the second according to high or low susceptibility of the subjects. Variation of these two independent variables might be diagrammed as in Figure 10.1. But since this was a

FIGURE 10.1.
Variation of two independent variables, each in two ways.

factorial design, all possible combinations of the values of the independent variables were used, as indicated in Table 10.1.

Table 10.1. *Diagram of a factorial design.*

		Degree of hypnosis	
		Hypnotized	Not hypnotized
Hypnotic Susceptibility	Low	(1)	(2)
	High	(3)	(4)

Table 10.1 shows that there are four possible combinations of the values of the independent variables. Each possible combination is represented by a square, a *cell*: (1) hypnotized and low susceptibility; (2) not hypnotized and low susceptibility; (3) hypnotized and high susceptibility; (4) not hypnotized and high susceptibility. With four experimental conditions there are four groups to consider in the experiment. Therefore, an equal number of subjects was assigned to each of the conditions.¹

More precisely, once the subjects were tested for hypnotic susceptibility, two classes of them were formed: those high and those low in susceptibility.

¹It is not necessary to have an equal number of subjects in each cell, but the statistical analysis is more complicated with unequal *n*'s.

Then those high in susceptibility were randomly assigned to either the hypnotic or the non-hypnotic conditions and similarly for those who tested out to be low in susceptibility.

The experiment was then conducted essentially as follows. First, all subjects, while in the waking state, were presented with a paired-associate learning task, and the number of errors that they required to learn the task was tabulated. A similar count was made on a comparable paired-associate list during the experimental conditions, and the dependent variable measure was the difference in number of errors made on the two occasions. The groups were treated as follows: Group 1 consisted of subjects for whom a test showed that they had low susceptibility to hypnosis, and they learned the second list while hypnotized; Group 2 was also made up of low-susceptibility subjects, but they learned the second list when in a normal awake state; Group 3 consisted of subjects who were quite susceptible to hypnosis, and they learned the second list while hypnotized; Group 4 was composed of highly susceptible subjects who learned the second list when not hypnotized. A statistical analysis of the dependent variable scores should then provide information concerning the following questions:

1. Does being hypnotized influence learning?
2. Does susceptibility to be hypnotized influence learning?
3. Is there an interaction between degree of hypnosis and susceptibility to be hypnotized?

The procedure for answering the first two questions is straightforward, but the third will require a little more consideration. Let us examine the dependent variable scores obtained for each group (Table 10.2)

Table 10.2. *Dependent Variable Scores for the Four Groups That Compose the Factorial Design of Table 10.1.*

GROUP			
1 (Hypnotized — low susceptibility)	2 (Not hypnotized — low susceptibility)	3 (Hypnotized — high susceptibility)	4 (Not hypnotized — high susceptibility)
0	9	-16	-4
-8	1	-20	8
1	-5	-20	-10
-20	-14	-41	9
-17	-2	-32	-10
-43	-3	-6	-23
-4	14	-42	29
-23	9	-29	-14
<i>n</i> : 8	8	8	8
ΣX : -114	9	-186	-15
ΣX^2 : 3148	593	6002	1927
\bar{X} : -14.25	1.12	-23.25	-1.88

Now let us place the means for the four groups in their appropriate cells (Table 10.3).

Table 10.3. *Showing Means for the Experimental Conditions.*

		Degree of hypnosis		
		Hypnotized	Not hypnotized	Means
Susceptibility	Low	-14.25	1.12	-6.57
	High	-23.25	-1.88	-12.57
Means:		-18.75	-.38	-9.57

Turning to the first question first, we shall study the effect of being in a hypnotized state on learning scores. For this purpose we shall ignore the susceptibility variable. That is, we have eight highly susceptible subjects who were hypnotized and eight subjects with low susceptibility who were hypnotized. Ignoring the fact that eight were high and eight were low in susceptibility, we have 16 subjects who learned while in a state of hypnosis. Similarly, we have 16 subjects who learned when they were not hypnotized. We therefore have two groups of subjects who, as a whole, were treated similarly except with regard to the hypnosis variable. For the hypnosis-nonhypnosis comparison it is irrelevant that half of each group were high in susceptibility and half were low in this respect — the susceptibility variable is balanced out. To make our comparison we need merely compute the mean for the 16 hypnotized subjects and for the 16 nonhypnotized subjects. To do this we have computed the mean of the means for the two groups of subjects who were hypnotized (Table 10.3). (This is possible because the n 's for each mean are equal.) That is, the mean of -14.25 and -23.25 is -18.75 and similarly for the nonhypnotized subjects, as shown in Table 10.3. Since the two means (-18.75 and $-.38$) are markedly different, we suspect that being hypnotized influenced the dependent variable. We shall, however, have to await the results of a statistical test to find out if this difference is significant.

Students who find it difficult to ignore the susceptibility variable when

considering the hypnosis variable should look at the factorial design as if they are conducting only one experiment and varying only the degree of hypnosis. In this case the susceptibility variable can be temporarily considered as an extraneous variable whose effect is balanced out. Thus, the two-groups design would look like that indicated in Table 10.4.

Table 10.4. *Looking at the Factorial Design as a Single Two-Groups Experiment.*

<i>Value of independent variable</i> →	<i>Group 1 (hypnotized)</i>	<i>Group 2 (Not hypnotized)</i>
<i>n</i>	16	16
Mean dependent variable score	-18.75	-.38

We now return to question number two and compare the high vs. low susceptibility classification by ignoring the hypnosis variable. The mean of the sixteen subjects who were low in susceptibility is -6.57 and the mean of the sixteen subjects who were high in susceptibility is -12.57. The difference between these means is not as great as before, suggesting that perhaps this variable did not greatly, if at all, influence the learning scores. Again, however, we must await the results of the test for significance before making a final judgment.

Now that we have preliminary answers to the first two questions, let us turn to the third. Is there an interaction between the two variables? *Interaction* is one of the most important concepts discussed in this book. If you adequately understand it, you will have ample opportunity to apply it in a wide variety of situations; it will shed light on a large number of problems and considerably increase your understanding of behavior.

First, let us approach the concept of interaction from an overly simplified point of view. Assume the problem is of the following sort: Is it more efficient (timewise) for a man who is dressing to put his shirt or his trousers on first? At first glance it might seem that a suitable empirical test would yield one of two answers: (1) shirt first or (2) trousers first. However, in addition to these possibilities there is a third answer — (3) it depends. Now “it depends” embodies the basic notion of interaction. Suppose a finer analysis of the data indicates what “it depends” on. We may find that it is more efficient for tall men to put their trousers on first but for short men to put their shirts on first. In this case we may say that our answer depends on the body build of the man who is dressing. Or to put it in terms of an interaction, we may say that there is an interaction between putting trousers or shirt on first with body build. This is the basic notion of interaction. Let us take another example from everyday life before we consider the concept in a more precise manner.

The author once had to obtain the support of a senior officer in the Army

to conduct an experiment. In order to control certain variables (e.g., the effect of the company commander) it was decided to use only one company. There were four methods of learning to be studied, so it was planned to divide the company into four groups. Each group (formed into a platoon) would then learn by a different method. The officer, however, objected to this design. He said that "we always train our men as a whole company. You are going to train the men in platoon sizes. Therefore, whatever results you obtain with regard to platoon-size training units may not be applicable to what we normally do with company-size units." The author had to admit this point, and it is quite a sophisticated one. It is possible that the results for platoons might be different than the results for companies — that there is an *interaction* between size of training unit and the type of method used. In other words, one method might be superior if used with platoons, but another if used with companies. Actually, previous evidence suggested that such an interaction was highly unlikely in this situation, so the author didn't worry about it; he only left a slightly distressed senior officer.

An interaction exists between two independent variables if the dependent variable value that results from one independent variable is determined by the specific value assumed by the other independent variable. To illustrate, momentarily assume that there is an interaction between the two variables of degree of hypnosis (hypnotized and nonhypnotized) and susceptibility to being hypnotized (high and low). The interaction would mean that the results (learning scores) for degree of hypnosis would depend upon the degree of susceptibility of the subject. Or, more precisely one might state the interaction as follows: whether or not being hypnotized affects amount learned depends on the degree of susceptibility of the subjects.

To enlarge on our understanding of the concept of an interaction, let us temporarily assume certain fictitious values for the hypnosis experiment, values that indicate a lack of an interaction (Figure 10.2). On the horizontal axis we have shown the two values of the susceptibility variable. The data points represent fictitious means of the four conditions: point number one is the mean for the low susceptibility hypnotized group; two is for the low susceptibility nonhypnotized group; three, the high susceptibility hypnotized group; and four, the high susceptibility not hypnotized group. The line that connects points one and three represents the performance of the hypnotized subjects, half who were low and half high in susceptibility. The line through points number two and four represents the performance of the non-hypnotized subjects. If these were real data, what would be the effects of the independent variables? First, variation of the degree of susceptibility would be said to not affect learning, for both lines are essentially horizontal. Second, the not hypnotized performed better than the hypnotized subjects (the "not hypnotized" line is higher than the "hypnotized" line). And third, the difference between the low susceptibility, hypnotized group and the low susceptibility, not hypnotized group (Difference A) is about the same as the difference

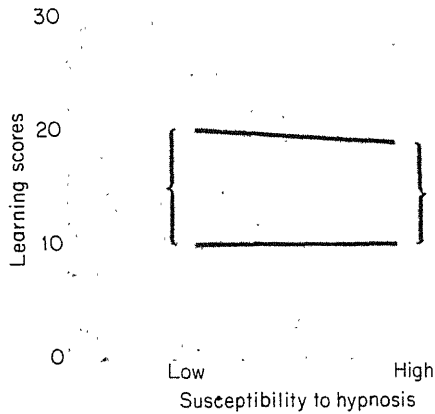


FIGURE 10.2.
Illustration of a lack of interaction with fictitious sample means.

between the high susceptibility, hypnotized and the high susceptibility, not hypnotized groups (Difference B). The performance of subjects who were and were not hypnotized is thus essentially independent of their degree of susceptibility. No interaction exists between these two variables. Put another way: if the lines drawn in Figure 10.2 are approximately parallel (i.e., if Difference A is approximately the same as Difference B), it is likely that no interaction exists between the variables.² However, if the lines based on these sample means are clearly not parallel (i.e., if Difference A is distinctly different from Difference B), an interaction is present.

Another way of illustrating the same point is to compute the differences between the means of the groups. The means plotted in Figure 10.2 are specified in the cells of Table 10.5. We have computed the necessary differences so that it can be seen that the difference between the subjects with low susceptibility who were hypnotized and who were not hypnotized is -10.00 and that for the high susceptibility subjects it is -8.75 . Since these are similar differences, there is probably no interaction present. The same conclusion would be reached by comparing differences in the other direction, i.e., since 0 and 1.25 are approximately the same, no interaction exists. Incidentally, the -10.00 is Difference A of Figure 10.2, and -8.75 is Difference B. Clearly if these differences are about the same, the lines will be approximately parallel.

At this point you may be disappointed that we did not illustrate an interaction. This can easily be arranged by assuming for the moment that the data came out as indicated in Table 10.6. In this case our lines would look

²Of course, as before, we are talking about sample values and not about population values. Thus, while this statement is true for sample values it is not true for population (true) values. Therefore, if the lines for the population values are even slightly nonparallel, there is an interaction.

Table 10.5. *Illustration of a Lack of an Interaction with Fictitious Means.*

		Degree of hypnosis		
		Hypnotized	Not hypnotized	Difference
Susceptibility	Low	10	20	-10.00
	High	10	18.75	-8.75
Difference		0.00	1.25	0.00

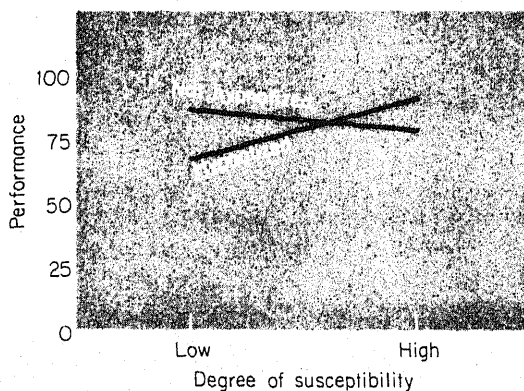
**FIGURE 10.3.**

Illustration of a possible interaction with fictitious sample means.

like those in Figure 10.3.

Now we note that the lines are not parallel; in fact they cross each other. Hence, if these were real data, we would make the following statements: Low susceptibility subjects who are not hypnotized are superior to low susceptibility subjects who are hypnotized; but high susceptibility subjects who are hypnotized are superior to high susceptibility subjects who are not hypnotized. Or, the logically equivalent statement is: The effect of being hypnotized depresses performance for low susceptibility subjects but facilitates performance for high susceptibility subjects. Put in other words: The difference between being hypnotized and being not hypnotized depends on the susceptibility of the subjects, or equally, the difference between degree of susceptibility depends on whether or not the subjects are hypnotized.

Table 10.6. *New Fictitious Means Designed to Show an Interaction.*

		Degree of hypnosis		
		Hypnotized	Not hypnotized	Means
Susceptibility	Low	69.1	90.0	79.55
	High	91.7	80.0	85.85
		80.40	85.00	82.70

This discussion should clarify the meaning of “interaction.” This is a rather difficult concept, however, and the examples in the remainder of the chapter should help to illuminate it further.

SUMMARY

When selected values of two or more independent variables are studied in all possible combinations, a factorial design is used. We have illustrated the factorial design by using two independent variables with two values of each. In this case subjects are assigned to the four experimental conditions. Analysis of the dependent variable data yields information on: (1) the influence of each independent variable on the dependent variable; and (2) the interaction between the two independent variables.

TYPES OF FACTORIAL DESIGNS

FACTORIAL DESIGNS WITH TWO INDEPENDENT VARIABLES

The 2 × 2 Factorial Design. The type of factorial design that we have discussed is referred to as the 2 × 2 factorial design. In this design we study the effect of two independent variables each varied in two ways. The number of numbers used in the label indicates the number of independent variables studied in the experiment. And the size of the numbers indicates the number of values of the independent variables. Since the 2 × 2 design has two numbers (2 and 2) we can tell immediately that there are two independent

variables. And since the numbers are both 2, we know that each independent variable assumed two values. From " 2×2 " we can also tell how many experimental conditions there are — 2 multiplied by 2 is 4.

The 3×2 Factorial Design. The 3×2 factorial design is one in which two independent variables are studied, one being varied in three ways, while the second assumes two values.

A 3×2 design was used in an experiment on programmed learning by Seidel and Rotberg (1966). The first variable concerned verbalizing the content of rules used in writing a computer program, and the three ways in which it was varied were: (1) Rules Condition — these subjects periodically wrote out the content of the rules; (2) Naming Condition — they wrote down the names of the rules; and (3) Nonverbalization Condition — they wrote their programs without any verbalization of the rules. The second variable was prompting vs. confirmation: (1) Prompting — under this condition subjects were required to write their answers after they were given the explicit information needed to make the response; (2) Confirmation — here subjects were required to write their answers prior to receiving information that confirmed the correctness of their responses. Subjects were then randomly assigned to these six conditions. A representation of this design is presented in Table 10.7.

Table 10.7. *A 3×2 Factorial Design.*

		Verbalization		
		Write rules	Name rules	None
Information	Prompting	17	34	368
	Confirmation	24	38	389

One set of dependent variable scores is, incidentally, included in Table 10.7. On the criterion tests the prompting vs. confirmation comparison was not significant. With regard to the verbalization variable, the subjects who did not have to verbalize anything regarding the rules (the nonverbalization condition) were superior to the subjects who served under the two verbalization conditions. Would you say that there is an interaction present?

The 3×3 Factorial Design. This design is one in which we investigate two independent variables, each varied in three ways. We therefore assign subjects to nine experimental conditions. Boe's (1966) use of a 3×3 design involved the effect of punishment duration and punishment intensity on the extinction of an instrumental response. Rats learned a barrier-crossing response and then were punished with shocks of various durations and intensities during extinction. More specifically, the three values of the intensity of the shock were .25 ma., 2.00 ma., and 4.00 ma. (ma. = milliamp). The durations of shock were .30 seconds, 1.00 seconds, and 3.00 seconds. The design is presented in Table 10.8, and five subjects were assigned to each cell. The dependent variable was a latency of response measure, but values for cells were not presented in the original article. In general, though, increases of both independent variables increased latency of responding (and therefore decreased response strength), as one would expect.

Table 10.8. *Illustration of a 3×3 Factorial Design.*

		Shock intensity		
		.25 ma.	2.00 ma.	4.00 ma.
Shock duration	.30 sec.			
	1.00 sec.			
	3.00 sec.			

The $K \times L$ Factorial Design. Each independent variable may be varied in any number of ways. The generalized factorial design for two independent variables may be labeled the $K \times L$ factorial design, where K stands for the first independent variable and its value indicates the number of ways in which it is varied; and L similarly denotes the second independent variable. K and L might then assume any value. If one independent variable is varied in four ways and the other in two ways, we would have a 4×2 design. If one independent variable is varied in six ways and the second in two ways, we would have a 6×2 design. If five values are assumed by one independent variable and three by the other we would have a 5×3 design, and so forth.

A 6×2 design was employed by Gampel (1966) in an experiment on verbal satiation (see p. 18). Verbal satiation for certain words is the loss of their meaning to a person when they are repeatedly presented. Briefly, this experimenter varied the duration of repetition of certain words in six ways: 0, 5, 15, 30, 60 and 120 seconds. The second variable was Word Category, and the two ways in which this variable was varied were: (1) use of the stimulus word that was "verbally satiated" or (2) use of a word that was strongly associated with that stimulus word. These two categories are referred to as "Stimulus Words" and "High Associate Words" respectively. The 6×2 factorial design is presented in Table 10.9.

Table 10.9. *Illustration of a $K \times L$ Factorial Design where $K = 6$ and $L = 2$.*

		Duration of repetition (Seconds)					
		0	5	15	30	60	120
Word category	Stimulus words						
	High associate words						

Gampel's dependent variable was defined as the amount of time required to find a stimulus word (or its high associate) within a large group of words. It was assumed that the longer the amount of time to search out and find a certain word, the greater the verbal satiation effect for that word. Her results showed that the greater the duration of repetition, the greater the search time (and hence the greater the verbal satiation). The effect of varying the second variable was that the high associate words required a longer search time than the stimulus words.

FACTORIAL DESIGNS WITH MORE THAN TWO INDEPENDENT VARIABLES

The $2 \times 2 \times 2$ Factorial Design. The previous factorial designs have concerned two independent variables. However, in principle the number of variables that can be studied is unlimited. The $2 \times 2 \times 2$ design is the simplest factorial for studying three independent variables. Following our

preceding rule this label implies that each of three independent variables is varied in two ways. It also follows that there are eight experimental conditions. As an illustration of a $2 \times 2 \times 2$ factorial design, consider an experiment by Tversky and Edwards (1966). These experimenters randomly assigned 24 undergraduate students to the eight different conditions and had each subject face a box that had two lights and three levers. Either the left or the right light would go on for a series of 1000 trials. By pressing the left or the right lever, the subject could guess which light was set to turn on, on the next trial. If he so guessed, no light flashed, but the response was automatically recorded. If the subject did not want to guess, he could depress the middle lever and observe which light came on. The three independent variables were: (1) instructional set; (2) probability value of the left vs. the right light coming on; and (3) response type. The two conditions of *instructional set* were: (1) nonstationary, in which case the subjects were told that the probability of each light coming on would change throughout the 1000 trials, or (2) stationary — the subjects were told that the probabilities remained constant. *Probability value* was either (1) 60:40 or (2) 70:30, i.e., for the former the left light came on 60 per cent of the time, as against 70 per cent for the latter. *Response type* conditions were: (1) Forced Prediction Group — these subjects *had* to predict which light was going to come on; (2) Free-Choice Group — these subjects did not have to make a prediction and could merely press the middle lever and observe which light was to come on. When the subject made a correct prediction he gained a nickel; otherwise he lost a nickel. This design is diagrammed in Table 10.10 and includes means for one of the dependent variable measures used.

Table 10.10. *Look vs. Bet: Average Number of Observation Trials for Each Experiment Group.*

Response type	Probability of left light			
	6		7	
	Instructional set		Instructional set	
	Stationary	Nonstationary	Stationary	Nonstationary
Free choice group	505	635	410	531
Forced prediction group	165	400	135	299

Briefly considering the results, we can see the following:

1. Response Type: Free Choice vs. Forced Prediction. The subjects who were free to observe the next trial (by pressing the middle lever) looked at those trials more than the subjects who were forced to give hypothetical predictions.
2. Instructional Set: Stationary vs. Nonstationary Instructions. Those subjects who were told that the probability might change during the sequence of 1000 trials looked more than those who received stationary instructions.
3. Probability Values: There was a tendency, though it was not significant, for the 60:40 group to look more than the 70:30 group.

The $K \times L \times M$ Factorial Design. It should now be apparent that any independent variable may be varied in any number of ways. The general case for the three independent variable factorial design is the $K \times L \times M$ design, where K , L , and M may assume whatever positive integer value the experimenter desires. For instance, if each independent variable assumes three values a $3 \times 3 \times 3$ design results. If one independent variable (K) is varied in two ways, the second (L) in three ways, and the third (M) in four ways, a $2 \times 3 \times 4$ design results. A $5 \times 3 \times 3$ design was used by Johnson and Bailey (1966) in an experiment on discrimination learning. Type of stimulus was varied in five ways, age of subject in three ways (kindergarten, fourth grade, or college) and number of stimulus-response bonds in three ways (one, two, or three). This design is diagrammed in Table 10.11.

Table 10.11. *Illustration of a $K \times L \times M$ Factorial Design where $K = 5$, $L = 3$ and $M = 3$.*

		Stimulus type				
		I	II	III	IV	V
One S-R bond	Age Kindergarten					
	4th grade					
	College					
Two S-R bonds	Age Kindergarten					
	4th grade					
	College					
Three S-R bonds	Age Kindergarten					
	4th grade					
	College					

STATISTICAL ANALYSIS OF FACTORIAL DESIGNS

We have compared the means for each of the experimental conditions in the hypnosis experiment and studied the concept of an interaction, but this has provided only tentative answers; firmer answers await the application of statistical tests to the data. Where we obtained a sizeable difference between the subjects who were hypnotized as against those who were not hypnotized for example, we said that we had to find out if the apparently sizeable difference in means was significant. The statistical analysis that is most frequently applied to the factorial design is *analysis of variance*, the rudiments of which were presented in Chapter 9. We shall limit our discussion to the 2×2 factorial design.

The first step in conducting an analysis of variance for the factorial design follows very closely that for any number of groups, as previously discussed. That is, we wish to compute the total *SS* and partition it into two major components, the among *SS* and the within *SS*. Let us return to the data in Table 10.2, which are summarized in Table 10.12.

Table 10.12. *A Summary of the Necessary Ingredients for Analysis of Variance.*
(Taken from Table 10.2.)

GROUP			
1	2	3	4
(Hypnotized — Low Susceptibility)	(Not Hypnotized — Low Susceptibility)	(Hypnotized — High Susceptibility)	(Not Hypnotized — High Susceptibility)
<i>n</i> : 8	8	8	8
ΣX : -114	9	-186	-15
ΣX^2 : 3148	593	6002	1927
\bar{X} : -14.25	1.12	-23.25	-1.88

To compute the total *SS*, we substitute the appropriate values from Table 10.12 in Equation (9.9) which for four groups (always the case for the 2×2 design) is:

$$\begin{aligned}
 (10.1) \quad \text{Total } SS &= (\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2) - \frac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4)^2}{N} \\
 &= (3148 + 593 + 6002 + 1927) - \frac{(-114 + 9 - 186 - 15)^2}{32} \\
 &= 8743.88
 \end{aligned}$$

Next, to compute the among groups SS , we substitute in Equation (9.10), which for four groups is:

$$(10.2) \quad \text{Among } SS = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} + \frac{(\sum X_4)^2}{n_4} - \frac{(\sum X_1 + \sum X_2 + \sum X_3 + \sum X_4)^2}{N}$$

$$\begin{aligned} \text{Among } SS &= \frac{(-114)^2}{8} + \frac{(9)^2}{8} + \frac{(-186)^2}{8} + \frac{(-15)^2}{8} - 2926.12 \\ &= 3061.12 \end{aligned}$$

And, as before, the within SS may be obtained by subtraction, Equation (9.11):

$$(10.3) \quad \begin{aligned} \text{Within } SS &= \text{total } SS - \text{among } SS \\ &= 8743.88 - 3061.12 = 5682.76 \end{aligned}$$

This completes the initial stage of the analysis of variance for a 2×2 factorial design, for we have now illustrated the computation of the total SS , the among SS , and the within SS . As you can see, the initial stage of the statistical computation is the same as that for the initial stage for a randomized groups design. But let us proceed farther.

The among groups SS tells us something about how all groups differ. However, we are interested not in simultaneously comparing all four groups, but only in certain comparisons. In a 2×2 factorial, of course, we have two independent variables, each varied in two ways. Hence we are interested in whether or not variation of each independent variable affects the dependent variable, and whether there is a significant interaction. The first step to perform is to compute the SS between groups for each independent variable. Using Table 10.1 (p. 246) as a guide, we may write our formulas for computing the between groups SS for the specific comparisons.

The groups are as labeled in the cells. Thus to determine whether or not there is a significant difference between the two values of the first (hypnosis) variable, we need to compute the SS between these two values. For this purpose we may use Equation (10.4):

$$(10.4) \quad \begin{aligned} &\text{SS between amounts of first independent variable} \\ &= \frac{(\sum X_1 + \sum X_3)^2}{n_1 + n_3} + \frac{(\sum X_2 + \sum X_4)^2}{n_2 + n_4} - \frac{(\sum X_1 + \sum X_2 + \sum X_3 + \sum X_4)^2}{N} \end{aligned}$$

For computing the SS between the conditions of the second independent variable we use Equation (10.5):

(10.5)

SS between amounts of second independent variable

$$= \frac{(\sum X_1 + \sum X_2)^2}{n_1 + n_2} + \frac{(\sum X_3 + \sum X_4)^2}{n_3 + n_4} - \frac{(\sum X_1 + \sum X_2 + \sum X_3 + \sum X_4)^2}{N}$$

Now, in our particular example we conduct tests to determine whether degree of hypnosis (our first independent variable) influences the dependent variable, whether hypnotic susceptibility (our second independent variable) influences the dependent variable, and whether there is a significant interaction. First, to determine the effect of being hypnotized we need to test the difference between the hypnotized and the nonhypnotized conditions. To make this test we shall ignore the susceptibility variable in the design.

Making the appropriate substitutions in Equation (10.4) we can compute the SS between the hypnosis conditions:

$$\begin{aligned} &= \frac{(-114 - 186)^2}{8 + 8} + \frac{(9 - 15)^2}{8 + 8} - 2926.12. \\ &= \frac{(90,000)}{16} + \frac{(36)}{16} - 2926.12 = 2701.13 \end{aligned}$$

This value will be used to answer the above question. However, we shall answer all questions at once, rather than piecemeal, so let us hold it until we complete this stage of the inquiry. We have computed a sum of squares among all four groups (i.e., 3061.12), and it can be separated into parts. We have computed one of these parts above, the sum of squares between the hypnosis conditions (2701.13). There are two other parts: the sum of squares between the susceptibility conditions, and the interaction. Let us compute the former, using Equation (10.5).

Substituting the required values in Equation (10.5) we determine that the "between" SS for the susceptibility conditions is:

SS between susceptibility conditions

$$= \frac{(-114 + 9)^2}{8 + 8} + \frac{(-186 - 15)^2}{8 + 8} - 2926.12 = 288.00$$

The among SS has three parts. We have directly computed the first two parts. Hence, the difference between the sum of the first two parts and the among SS provides the third part, in this case that for the interaction:

(10.6)

Interaction $SS =$ among $SS -$ between SS for first variable (hypnosis)
 $-$ between SS for second variable (susceptibility)

Recalling that the among SS was 3061.12, the between SS for the hypnosis conditions was 2701.13, and the between SS for the susceptibility conditions was 288.00, we find that the SS for the interaction is:

$$\text{Interaction } SS = 3061.12 - 2701.13 - 288.00 = 71.99$$

This completes the computation of the sums of squares. Let us add that these values should all be positive. If your computations yield a negative SS , check your work until you discover the error. There are only several minor matters to discuss before the analysis is completed. Before we continue, however, let us summarize our findings to this point in Table 10.13.

Table 10.13. *Sums of Squares for the 2×2 Factorial Design.*

<i>Source of Variation</i>	<i>Sum of Squares</i>
Among Groups	(3061.12)
Between Hypnosis (H)	2701.13
Between Susceptibility (S)	288.00
Interaction: hypnosis \times susceptibility	71.99
Within groups	5682.76
Total	8743.88

We now must discuss how to determine various degrees of freedom for this application of the analysis-of-variance procedure. Repeating the equations in Chapter 9, for the major components:

$$(10.7) \quad \text{Total } df = N - 1$$

$$(10.8) \quad \text{Among (or Between) } df = r - 1$$

$$(10.9) \quad \text{Within } df = N - r$$

In our example, $N = 32$, and r (number of groups) = 4. Hence, the total df is $32 - 1 = 31$, the among df is $4 - 1 = 3$ (the among df is based on four separate groups or conditions), and the within df is $32 - 4 = 28$. The similarity between the manner in which we partition the total SS and the total df may also be continued for the among SS and the among df . The among df is 3. Since we analyzed the among SS into three parts, we may do the same for the among df , one df for each part (one df for each part is only true for a 2×2 factorial design). Take the hypnosis conditions first. Since we are temporarily ignoring the susceptibility variable, we have only two conditions of hypnosis to consider, or, if you will, two groups. Hence, the df for the between-hypnosis conditions is based on $r = 2$. Substituting this value in Equation (10.8), we see that the between-hypnosis df is $2 - 1 = 1$. The same holds true for the susceptibility variable; there are two values, hence $r = 2$ and the df for this source of variation is $2 - 1 = 1$.

Now for the interaction df . Note in Table 10.14 that the interaction is written as hypnosis \times susceptibility. We may, of course, abbreviate the notation, as is frequently done, by using $H \times S$. This is read "the interaction between hypnosis and susceptibility." The " \times " sign may be used as a mnemonic device for remembering how to compute the interaction df : multiply the number of degrees of freedom for the first variable by that for the second. Since both variables have one df , the interaction df is also one, i.e., $1 \times 1 = 1$. This accounts for all three df that are associated with the among SS .³ These findings are added to Table 10.13, forming Table 10.14.

Table 10.14. *Sums of Squares and df for the 2×2 Factorial Design.*

<i>Sources of Variation</i>	<i>Sums of Squares</i>	<i>df</i>
Among groups	(3061.12)	(3)
Between hypnosis (H)	2701.13	1
Between susceptibility (S)	288.00	1
Interaction: hypnosis \times susceptibility (H \times S)	71.94	1
Within groups	5682.76	28
Total	8743.88	31

In the 2×2 factorial design there are four mean squares in which we are interested. In this experiment they are: between hypnosis conditions, between susceptibility conditions, the interaction, and within groups. To compute the mean square for the between hypnosis source of variation, we divide that sum of squares by the corresponding df :

$$\frac{2701.13}{1} = 2701.13$$

Similarly the within-groups mean square is computed:

$$\frac{5682.76}{28} = 202.95$$

These values are then added to our summary table of the analysis of variance, as we shall show shortly.

This completes the analysis of variance for the 2×2 design, at least in the usual form. We have analyzed the total sum of squares into its various components. In particular, we have several sources of between sums of squares to study and a term that represents the experimental error (the within-groups mean square). We said that the "between" components

³If this is not clear to you, then you might merely remember that the df for the between SS in a 2×2 design are always the same, as shown in Table 10.14. That is, the df for the among SS is always 3, the df for the SS between each independent variable condition is 1, and for the interaction, 1.

indicate the extent to which the various experimental conditions differ. For instance, if any given "between" component, such as that for the hypnosis conditions, is sizeable, then that may be taken to indicate that hypnosis influences the dependent variable. Hence we need merely conduct the appropriate F -tests to determine whether or not the various "between" components are significantly larger than would be expected by chance. The first F for us to compute is that between the two conditions of hypnosis.⁴ To do this we merely substitute the appropriate values in Equation (9.17). Since the mean square between the hypnosis conditions is 2701.13 and the mean square within groups is 202.95, we divide the former by the latter:

$$F = \frac{2701.13}{202.95} = 13.30$$

The F between the hypnosis susceptibility conditions is:

$$F = \frac{288.00}{202.95} = 1.41$$

And the F for the interaction is:

$$F = \frac{71.94}{202.95} = .35$$

These values have been entered in Table 10.15.

Table 10.15. *Complete Analysis of Variance of the Performance Scores.*

Source of Variation	Sum of Squares	df	Mean Square	F
Between hypnosis	2701.13	1	2701.13	13.30
Between susceptibility	288.00	1	298.00	1.41
Interaction: H \times S	71.99	1	71.99	.35
Within groups (error)	5682.76	28	202.95	
Total	8743.88	31		

Following the preceding discussion, let us observe that Table 10.15 is the final summary of our statistical analysis. This is the table that should be presented in the results section of an experimental write-up. We next assign a probability level to these values. That is, we need to determine the odds

⁴The factorial design offers us a good example of a point we made in Chapter 9. That is that if we are specifically interested in certain questions, then there is no need to conduct an F -test for the among groups source of variation. With this design we are exclusively interested in whether our two independent variables are effective and whether there is an interaction. Hence we proceed directly to these questions without running an overall F -test among all four groups, although such may be easily conducted.

that the F 's could have occurred by chance. As before we first set up a null hypothesis: There is no difference between our groups. However, we have three more precise null hypotheses in this type of design:

1. There is no difference between the means of the two conditions of hypnosis.
2. There is no difference between the means of the two degrees of hypnotic susceptibility.
3. There is no interaction between the two independent variables.⁵

The general strategy is to determine the probability associated with each value of F . Assuming that we have set a significance level of 0.05 for each F -test, we need merely determine the probability associated with each F . If that probability is 0.05 or less, we can reject the appropriate null hypothesis and conclude that the independent variable in question was effective in producing the result.⁶

Let us turn to the first null hypothesis, that for the hypnosis variable. Our obtained F is 13.30. We have one df for the numerator and 28 df for the denominator. We can determine that an F of 4.20 is required for significance at the 5 per cent level with 1 and 28 df (Table 9.11). Since our F of 13.30 exceeds this value, we may reject the first null hypothesis. The conclusion is that the two conditions of hypnosis led to significantly different performance. And since the mean for the hypnosis condition (-18.75) is lower than for the nonhypnosis condition ($-.38$), the authors concluded that hypnosis has "a strong inhibiting effect on learning" (Harley and Harley, 1966, p.1).

To test the effect of varying hypnotic susceptibility, we note that the F ratio for this source of variation is 1.41. We have 1 and 28 df available for this test. The necessary F value is, as before, 4.20. Since 1.41 does not exceed 4.20, we conclude that variation of hypnotic susceptibility does not significantly influence amount learned.

To study the interaction, refer to Figure 10.4. Note that the lines do not deviate to any great extent from being parallel, suggesting that there is no significant interaction between the variables. Incidentally, the fact that the line for the nonhypnotized condition is noticeably higher than that for the hypnotized condition is a graphic illustration of the effectiveness of the hypnosis variable.

To test the interaction we note that the F is .35. This F is considerably below 1.00. We can, therefore, conclude immediately that the interaction is

⁵A more precise statement could be made: There is no difference in the means of the four groups after the cell means have been adjusted for row and column effects. However, such a statement probably will only be comprehensible to you after further work in statistics.

⁶Of course, assuming adequate control procedures have been exercised.

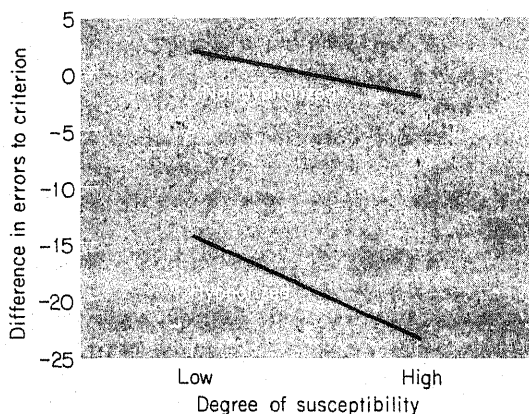


FIGURE 10.4.

The actual data of Harley and Harley (1966) suggest that there is a lack of interaction between hypnotic susceptibility and degree of hypnosis.

not significant. A check on this may be made by noting that we also have 1 and 28 *df* for this source of variation. And we know that an *F* of 4.20 is required for significance at the 5 per cent level. Clearly .35 does not approach 4.20, and hence is not significant. The third null hypothesis is not rejected.

The preceding discussion for the statistical analysis of a factorial design has been rather lengthy because of its detailed nature. Now with this background it is possible for us to breeze through the next example. This experiment, reported by Lachman, Meehan, and Bradley (1965), was an investigation of the effect of two independent variables on the learning of concepts.⁷ The general procedure required the subjects to face a box on which two complex stimulus pictures were presented. The pictures had borders that were black, white, or red, and within the borders were geometrical forms (circles and squares) that also varied in color. The subject's task was to guess which of the two pictures was correct by selecting a color on one of them. When he was right, a light at the top of the box came on. The pairs of stimulus pictures were changed on each trial, and the subject continued making choices until he found the color that was always on one picture, as indicated by his making ten successive correct choices. An example of a correct concept would be always choosing the picture that had a red border. Once this initial learning was complete, a second set of stimuli was presented, and the above procedure was repeated as the subjects solved their second problem.

⁷The very interesting theoretical questions that led to this experiment concerned the mechanisms for verbal mediation. The reader is referred to the original article for a discussion of the more theoretical issues.

The general question that Lachman *et al.* studied concerned the relationships between the correct color on the initial learning task to the color that was correct during the later learning period. There were two such relationships. The first was the strength of the association between the names of the colors that were correct on the first and the second tasks. We may, for instance, note that the words "black" and "white" are more strongly associated than are the words "red" and "white." So we have two degrees of word association: high and low, and half of the subjects served under each condition. For example, the high word association subjects might have a black symbol as correct during the first learning period, and a white symbol as correct during the second. The low word association subjects would first have a red symbol as the correct one, and a white symbol correct on the second task. The question, then, is: Does variation of the strength of the word association between critical symbols influence the rapidity of learning a concept?

The second question concerned the observing response. Roughly, an observing response is varied by varying the location of the critical color on the stimulus pictures. It was varied in two ways: (1) held constant from the first to the second learning periods, or (2) it was changed. An example of the constant observing response condition would be that if a red border of the stimulus picture was correct during initial learning, the color of the border was also correct during later learning (though the color of the border would be changed). But for the changed observing response conditions, if a red border was first correct, then the color of one of the geometrical forms within the border was correct during the second learning period. In short, the second question was whether varying the observing response (by changing the location of the critical color on the stimulus pictures) influences the rapidity of learning a new concept. The third question, of course, is whether there is an interaction between the word association variable and the observing response variable.

A diagram of the 2×2 factorial design is presented in Table 10.16, and the three null hypotheses that were tested are as follows:

1. There is no difference between the means for the high and low word association conditions.
2. There is no difference between the means for the observing response conditions.
3. There is no interaction between the word association and the observing response variables.

Twelve subjects were randomly assigned to each cell, and the number of trials to reach criterion are presented for each subject in Table 10.17.

Table 10.16. *A 2 × 2 Factorial Design with Strength of Word Association and Observing Response as the Two Variables.*

		Word association	
		High	Low
Observing response	Changed	17.08	32.25
	Constant	6.42	16.25

Table 10.17. *Number of Trials to Criterion.*

GROUP			
1 <i>Changed OR — High Association</i>	2 <i>Changed OR — Low Association</i>	3 <i>Constant OR — High Association</i>	4 <i>Constant OR — Low Association</i>
23	12	3	33
10	3	1	2
4	32	1	2
10	18	5	1
34	12	75	1
14	10	75	4
15	17	5	5
31	28	2	12
75	59	2	4
75	4	19	10
75	6	5	2
21	4	2	1
<hr/>			
$\Sigma X:$ 387	205	195	77
$\Sigma X^2:$ 20599	6367	11709	1405
$\bar{X}:$ 17.08	32.25	6.42	16.25

Our first step will be to compute the total SS by substituting the values in Table 10.17 in Equation (10.1):

$$\begin{aligned}
 \text{Total } SS &= 20599 + 6367 + 11709 + 1405 - \frac{(387 + 205 + 195 + 77)^2}{48} \\
 &= 24,528.00
 \end{aligned}$$

Now we shall compute the among SS by appropriate substitutions in Equation (10.2).

$$\begin{aligned}\text{Among } SS &= \frac{(387)^2}{12} + \frac{(205)^2}{12} + \frac{(195)^2}{12} + \frac{(77)^2}{12} - \frac{(387 + 205 + 195 + 77)^2}{48} \\ &= 4093.60\end{aligned}$$

The within SS is [see Equation (10.3)]:

$$\text{Within } SS = 24,528.00 - 4093.60 = 20,434.40$$

As our next step we shall analyze the among SS into its three components: between the word association condition, the observing response condition, and the $WA \times OR$ interaction. Considering word association first, we substitute the appropriate values in Equation (10.4) and find that:

SS between word association conditions

$$\begin{aligned}&= \frac{(387 + 195)^2}{12 + 12} + \frac{(205 + 77)^2}{12 + 12} - \frac{(387 + 205 + 195 + 77)^2}{48} \\ &= 1875.00\end{aligned}$$

Substituting in Equation (10.5) to compute the SS between the two conditions of observing response:

SS between observing response conditions

$$\begin{aligned}&= \frac{(387 + 205)^2}{12 + 12} + \frac{(195 + 77)^2}{12 + 12} - \frac{(387 + 205 + 195 + 77)^2}{48} \\ &= 2133.34\end{aligned}$$

The SS for the interaction component may now be seen to be:

$$4093.60 - 1875.00 - 2133.24 = 85.26$$

The various df may now be determined.

$$\begin{aligned}\text{Total } (N - 1) &= 48 - 1 = 47 \\ \text{Over-all between } (r - 1) &= 4 - 1 = 3 \\ \text{Between word association} &= 2 - 1 = 1 \\ \text{Between observing response} &= 2 - 1 = 1 \\ \text{Interaction: } WA \times OR &= 1 \times 1 = 1 \\ \text{Within } (N - r) &= 48 - 4 = 44\end{aligned}$$

The mean squares and the F 's have been computed and placed in the summary table (Table 10.18).

Table 10.18. *Summary of the Analysis of Variance for the Concept Learning Experiment.*

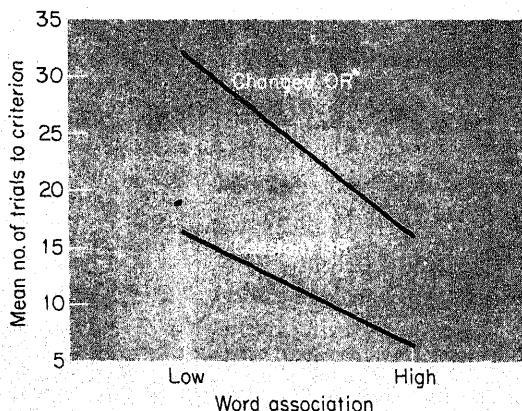
Source of Variation	Sum of Squares	df	Mean Square	F
Between word association	1875.00	1	1875.00	4.04
Between observing response	2133.34	1	2133.34	4.59
Interactions WA \times OR	85.36	1	85.36	.18
Within groups	20,434.40	44	464.42	
Total	24,528.00	47		

INTERPRETING THE F 's

To test the F for the word association variable, let us note that we have 1 and 44 degrees of freedom available. Assuming a .05 level test, we enter Table 9.11 and find that we must interpolate between 40 df and 60 df . The F values are 4.08 and 4.00 respectively. Consequently, an F with 1 and 44 df must exceed 4.06. The F for word association is 4.04; we therefore fail to reject the first null hypothesis and must conclude that variation of the word association variable did not significantly affect rapidity of concept learning.

We have the same number of df available for evaluating the effect of the observing response variable, and therefore the F for this effect must also exceed 4.06 in order to be significant. We note that it is 4.59, and we can thus reject the second null hypothesis. The empirical conclusion is that variation of the observing response significantly influences rapidity of forming a concept.

Referring to Figure 10.5 we can visually study these findings. First, observe that the data points are lower for the high word association condition

**FIGURE 10.5.**

Data points for the concept learning experiment. Since the lines are approximately parallel, there is a lack of interaction.

than for the low word association condition, but this decrease is not significant. The points for the changed observing response conditions are higher than for the constant observing response condition. Since this variable significantly influenced the dependent variable scores, maintaining a constant observing response facilitated the formation of a concept.

Finally, we note that the two lines are approximately parallel. The suggestion is thus that there is a lack of interaction between the independent variables, a suggestion that is confirmed by the F value for the interaction source of variation, viz., this F is well below one and we can thus immediately conclude that it is not significant.

This completes our examples of the statistical analysis of factorial designs. We have discussed factorial designs generally, but have only illustrated the analysis for the 2×2 case. If you are interested in obtaining general principles for the analysis of any factorial design you should consult one of the references previously given or plan on taking a more advanced course. It is not likely, however, that you will get beyond the 2×2 design in your elementary work.

THE CHOICE OF A CORRECT ERROR TERM

One of the most important problems of statistical analysis is the choice of the correct error term. With reference to the F -test the problem is one of choosing the correct denominator. The error term that we have used is the within-groups mean square. Although the within-groups mean square is usually the correct error term, we should be aware that sometimes it is not. To understand this, let us note that there are three types of factorial designs that you are likely to encounter in your work. The first is the case of a fixed model, the second is the case of a random model, and the third is the case of a mixed model. ("Model" in each case refers to characteristics of the independent variables of the factorial design.) Let us take these three cases in turn and illustrate them by means of a 2×2 design.

THE CASE OF A FIXED MODEL

The 2×2 design indicates that we have two independent variables, each varied in two ways. Now, if we have some particular reason to select the two values of the two variables, it can be said that we are dealing with a fixed model. This is so because we have not arrived at the particular values of the independent variables in a random manner. In other words, we have chosen the two values of each independent variable in a premeditated way. We are

interested in method A of teaching (a specific method) versus method B, for example. Or we choose to study 10 hours of training versus 20 hours. Similarly, we decide to give our rats 50 versus 100 trials, selecting these particular values for a special reason. When we select our values of the independent variables for some specific reason, and do not select them at random, we have the case of a fixed model. *For this case the within-groups mean square is the correct error term for all F-tests being run.* If we refer to our two independent variables as K and L, and the interaction as $K \times L$, we have the following between-groups mean squares to test: that between the two conditions of K, that between the two conditions of L, and that for $K \times L$. For a fixed model, each of the between-groups mean squares should be divided by the within-groups mean square. This is, incidentally, the case most frequently encountered in psychological research.

THE CASE OF A RANDOM MODEL

If the values of the two independent variables have been selected at random, you are using a random model. For example, if our two variables are number of trials and IQ of subjects, we would consider all possible reasonable numbers of trials and all possible reasonable IQ's. Our two particular values of each independent variable would then be selected strictly at random. For instance, we might consider as reasonable possible values of the first independent variable — numbers of trials — those from 6 to 300. We would then place these 295 numbers in a hat and draw two from them. The resulting numbers would be the values that we would assign to our independent variable. The same process would be followed with regard to the IQ variable. In the case of studying various characteristics of subjects such as IQ, however, we would probably do the following. If a random sample of subjects has been drawn, then merely grouping subjects into classes would satisfy our requirement. That is, we might divide all subjects into two groups, using a certain IQ score as the dividing line. This would constitute random values of this independent variable. The reason that this is so is that we have specified that our subjects were selected at random from a given population. Hence, in randomly selecting subjects, we also randomly select values of their characteristics, such as IQ.

The procedure for testing the between-groups mean squares for the case where both independent variables are random variables is as follows: Test the interaction mean square by dividing it by the within-groups mean square. Then test the other mean squares by dividing them by the interaction mean square. That is, test the $K \times L$ mean square by dividing it by the mean square within groups. Then test the mean square between the two conditions of K by dividing it by the $K \times L$ mean square, and also test the mean square between L by dividing it by the

$K \times L$ mean square. We might remark that designs in which both variables are random are relatively rare in psychological research.

THE CASE OF A MIXED MODEL

This is a less uncommon case than that where both variables are random, but still does not occur as frequently as the case of a fixed model. The case of a mixed model occurs when one independent variable is fixed and the other is random. The procedure for testing the three mean squares for this case is as follows: *Divide the within-groups mean square into the interaction mean square; divide the interaction mean square into the mean square for the fixed independent variable; and divide the within mean square into the mean square for the random independent variable.*

These are the three cases that are most likely to be encountered in your research, though there are a number of variations that can occur.⁸ The importance of these rules may not be immediately apparent, but a consideration of the topic of generalization (Chapter 14) will correct that. Furthermore, in the appendix to this chapter we briefly present a rationale that will help you to remember these rules and to understand the reasons for them. See the excellent book by Wine (1964) for an elaboration.

THE IMPORTANCE OF INTERACTIONS

Our goal in psychology is to arrive at statements about the behavior of organisms that can be used for such purposes as explaining, predicting, and controlling behavior. For accomplishing these purposes we would like our statements to be as simple as possible. Behavior is anything but simple, however, and it would be very surprising to us if our *statements* about behavior were simple. It would seem more reasonable to expect that complex statements must be made about complex events. Those who talk about behavior in simple terms are likely to be wrong. This is illustrated by "common sense" discussions of behavior. People often say such things as "he is smart, he will do well in college," or "she is pretty, she will have no trouble finding a husband." However, such matters are not that uncomplicated; there are variables other than "smartness" that influence how well a person does in college, and there are variables other than beauty that influence a girl's marriageability. Furthermore, such variables do not always act on all

⁸For a more thorough consideration of this topic you are referred to Anderson and Bancroft (1952, Chapter 23), or to Wine (1964); if you prefer a psychological text, see Lindquist (1953) or Winer (1962).

people in the same manner. Rather, they *interact* in such a way that people with certain characteristics behave one way, but people with the same characteristics in addition to other characteristics behave another way. Let us illustrate by speculating about two variables that might influence the likelihood of a girl's getting married: beauty and intelligence. Consider two values of each of these variables: beautiful and not beautiful; high intelligence and low intelligence. Were we to study these variables, we would collect data on a sample of four groups of girls: beautiful girls with high intelligence, beautiful girls with low intelligence, not-beautiful girls with high intelligence and not-beautiful girls with low intelligence. Suppose our dependent variable to be the frequency with which girls in these four groups married, and we found that: beautiful girls with low intelligence get married most frequently, not-beautiful girls with high intelligence get married next most frequently, beautiful girls with high intelligence get married with the third greatest frequency, and low-intelligence girls who are not beautiful get married the least frequently (see Figure 10.6).

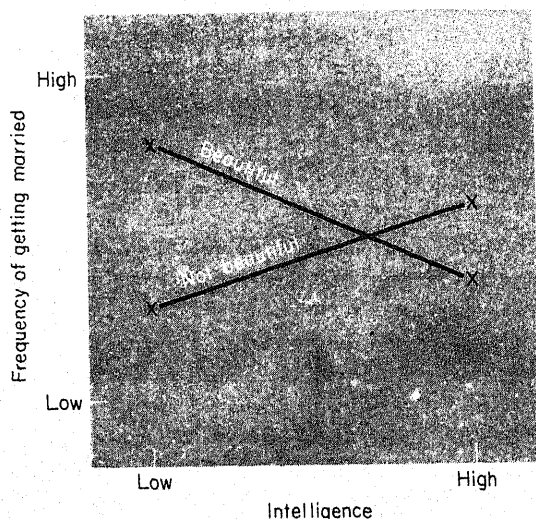


FIGURE 10.6.

A possible interaction between beauty and intelligence.

Now, if these findings were actually obtained, then the simple statement, "she is pretty, she will have no trouble finding a husband" is inaccurate. Beauty is not the whole story; intelligence is also important. We cannot say that beautiful girls are more likely to get husbands any more than we can say that unintelligent girls are more likely to get husbands. The only accurate

statement is that beauty and intelligence interact; beautiful girls with low intelligence are more likely to get married than unbeautiful girls with low intelligence; but unbeautiful girls with high intelligence are more likely to get married than beautiful girls with high intelligence. Incidentally it could be argued that these fictitious findings are reasonable. For clearly, boys like girls to be beautiful, but we often hear (perhaps erroneously) that "boys don't like girls to be too intelligent." Hence the superiority of low intelligence beautiful girls. High intelligence girls who are not beautiful might be next, for they are smart enough to compensate for their lack of beauty, in addition to the possibility that high intelligence girls who are beautiful might just "scare" boys away with their "dazzling" combination. In any event, the answer to these questions must await suitable research.

We have just barely begun to make completely accurate statements when we talk about interactions between two variables. It is highly likely that interactions of a much higher order occur, that is, interactions among three, four, or any number of variables. To illustrate, not only might beauty and intelligence interact, but in addition such variables as desire to get married (too high a desire might get the boys "worried" too soon), economic status of the parents, social graces, and so on. Hence, for a really adequate understanding of behavior, we need to determine the large number of interactions that undoubtedly occur. In the final analysis, if such ever occurs in psychology, we will probably arrive at two general kinds of statements about behavior: those statements that tell us how everybody behaves (those ways in which people are similar), with no real exceptions; and those statements that tell us how people differ. The latter will probably involve statements about interactions. For people with certain characteristics act differently than people with other characteristics in the presence of the same stimuli. And statements that describe the varying behavior of people will probably rest on accurate determination of interactions. If such a complete determination of interactions ever comes about, we will be able to understand the behavior of what humanists call the "unique" personality.

Now let us refer the concept of interaction back to a topic discussed in Chapter 2. We discussed ways in which we become aware of a problem, one of them being as a result of contradictory findings in a series of experiments. For example, we considered two experiments, each using the same design and performed on the same problem, but with contradictory results. Why? One reason might be that a certain variable was not controlled in either experiment. Hence, it might assume one value in the first experiment and a second value in the second experiment. And if such an extraneous variable interacts with the independent variable(s), then the discrepant results become understandable. A new experiment could then be conducted in which that extraneous variable becomes an independent variable. As it is purposefully manipulated along with the original independent variable, the na-

ture of the interaction can be determined. In this way not only will the apparently contradictory results be understood, but a new advance in knowledge will be made.

This situation need not be limited to the case where the extraneous variable is uncontrolled. For instance, the first experimenter may hold the extraneous variable constant at a certain value while the second experimenter may also hold it constant, but at a different value. And the same result would obtain as when the variable went uncontrolled — contradictory findings in the two experiments. Let us illustrate by returning to the experiment on language suppression that was discussed on pp. 18–19. Recall that the first investigator found that the experimental treatment successfully produced a suppression effect for a pronoun for the experimental group but there was no suppression for the control group. The relevant extraneous variable was the location of the experimenter, and in this study the subjects could not see the experimenter. In the repetition of the experiment, however, the subject *could* see the experimenter, and the results were that there was no pronoun suppression for the experimental, as compared with the control group. The ideal solution for this problem, we said, would come by conducting a new experiment using a factorial design that incorporates experimenter location as the second variable. Hence, as shown in Table 10.19, the first variable is the original one (prior verbal stimulation), and we have varied it in two ways by using an experimental and a control group. The second variable, experimenter location, has two values: the subject cannot see the experimenter, and the subject can see him.

In short, we are repeating the original experiment, under two conditions of the extraneous variable. A graphic illustration of the expected results is

Table 10.19. *A Design to Investigate Systematically the Effect of an Extraneous Variable.*

		Prior verbal stimulation	
		None (control group)	Some (experimental group)
Location of experimenter	Subject cannot see him		
	Subject can see him		

offered in Figure 10.7. We can see there that the experimental group exhibits a larger suppression effect than does the control group when the subjects cannot see the experimenter. But where the subjects *can* see the experimenter, there is no significant difference between the two groups. It is determined

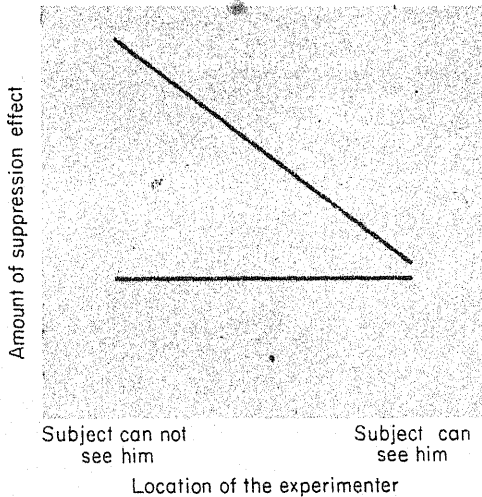


FIGURE 10.7.

Illustration of an interaction between the independent variable and location of the experimenter. When the experimenter's location was systematically varied, the reason for conflicting results in two experiments became clear.

that there is an interaction between the location of the experimenter and the variable of prior verbal stimulation. What at first looked like a contradiction is resolved by isolating an interaction between the original independent variable and an extraneous variable. The problem is solved by resorting to a factorial design.

Undoubtedly these considerations hold for a wide variety of experimental findings, for the contradictions in the psychological literature are legion. By shrewd application of factorial designs to such problems their resolution should be accomplished.

VALUE OF THE FACTORIAL DESIGN

Not long ago the standard design in psychological research was the two-groups design. For many years, however, statisticians and researchers in such fields as agriculture and genetics, had been developing other kinds of designs. One of these was the factorial design, which, incidentally, grew with the development of analysis of variance. Slowly, psychologists started trying out

these designs on their own problems. Some of them were found to be inappropriate, but the factorial design is one that has enjoyed success, and the extent of its success is still widening. As but one example, Kiesler (1966), in discussing research in the area of clinical psychology, says that "... psychotherapy research can no longer ignore the necessity for factorial designs" (Kiesler, 1966, p. 133). It is particularly applicable to psychological problems, although some psychologists still hold that the two-groups design should be the standard for our kinds of research. Our position has been that each type of design that we have considered is appropriate for particular situations. We cannot say that one design should *always* be used and that the others are useless. However, we would make the general statement that where it is feasible, the factorial design is superior to the other designs that we discussed. Professor Ronald Fisher has made some interesting comments on this point.

In expositions of the scientific use of experimentation it is frequent to find an excessive stress laid on the importance of varying the essential conditions *only one at a time*. The experimenter interested in the causes which contribute to a certain effect is supposed, by a process of abstraction, to isolate these causes into a number of elementary ingredients, or factors, and it is often supposed, at least for purposes of exposition, that to establish controlled conditions in which all of these factors except one can be held constant, and then to study the effects of this single factor, is the essentially scientific approach to an experimental investigation. This ideal doctrine seems to be more nearly related to expositions of elementary physical theory than to laboratory practice in any branch of research. In experiments merely designed to illustrate or demonstrate simple laws, connecting cause and effect, the relationships of which with the laws relating to other causes are already known, it provides a means by which the student may apprehend the relationship, with which he is to familiarize himself, in as simple a manner as possible. By contrast, in the state of knowledge or ignorance in which genuine research, intended to advance knowledge, has to be carried on, this simple formula is not very helpful. We are usually ignorant which, out of innumerable possible factors, may prove ultimately to be the most important, though we may have strong presuppositions that some few of them are particularly worthy of study. We have usually no knowledge that any one factor will exert its effects independently of all others that can be varied, or that its effects are particularly simply related to variations in these other factors. On the contrary, when factors are chosen for investigation, it is not because we anticipate that the laws of nature can be expressed with any particular simplicity in terms of these variables, but because they are variables which can be controlled or measured with comparative ease. *If the investigator, in these circumstances, confines his attention to any single factor, we may infer either that he is the unfortunate victim of a doctrinaire theory as to how experimentation should proceed, or that the time, material or equipment at his disposal is too limited to allow him to give attention to more than one narrow aspect of his problem.* The modifications possible to any complicated apparatus, machine or industrial process must always be considered as potentially interacting with one another, and must be judged by the probable effects of such interactions. If they have to be tested one at a time this is not because to do so is an ideal scientific procedure, but because to test them simultaneously would sometimes be too

troublesome, or too costly. In many instances, the belief that this is so has little foundation. Indeed, in a wide class of cases an experimental investigation, at the same time as it is made more comprehensive, may also be made more efficient if by more efficient we mean that more knowledge and a higher degree of precision are obtainable by the same number of observations. (Fisher, 1953, pp. 91-92. *Italics ours.*)

Let us look into this matter more thoroughly. First we may note that the amount of information obtained from a factorial design is considerably greater than that obtained from the other designs mentioned, relative to the number of subjects used. For example, let us say that we have two problems: (1) does variation of independent variable K affect a given dependent variable; and (2) does variation of independent variable L affect that same dependent variable. If we investigated these two problems by the use of a two-groups design, we would obtain two values for each variable. That is, K will be varied in two ways (K_1 and K_2), and similarly for L (L_1 and L_2). With 60 subjects available for each experiment, the design for the first problem would be:

<i>Experiment #1</i>	
Group K_1	Group K_2
30 subjects	30 subjects

And similarly for the second problem:

<i>Experiment #2</i>	
Group L_1	Group L_2
30 subjects	30 subjects

With a total of 120 subjects we are able to evaluate the effect of the two independent variables. We would not be able to tell if there is an interaction between K and L if we looked at these as two separate experiments.

But what if we used a factorial design to answer our two problems? Assume that we will want 30 subjects for each condition. In this case the factorial would be as in Table 10.20. We would have four groups with 15 subjects per group. But for comparing the two conditions of K we would have 30 subjects for condition K_1 and 30 subjects for K_2 . This is just what we had for Experiment #1. And the same for the second experiment: We have 30 subjects for each condition of L. Here we have accomplished everything with the 2×2 factorial design that we would have accomplished with the two separate experiments with two groups. But with those two experiments we required 120 subjects in order to have 30 available for each condition; however, with the factorial design we need only 60 subjects to have the same number of subjects for each condition. The factorial design is much more efficient because we use our subjects simultaneously for testing both independent variables.⁹ In addition, we can evaluate the interaction between K and

⁹We are assuming that the two two-groups experiments are analyzed independently. In a very atypical case, the error terms might be pooled, which would result in a larger number of degrees of freedom than for the factorial design.

Table 10.20. *A 2×2 Design that Incorporates Two, Two-Groups Experiments. The numbers of subjects for cells, conditions, and the total number in the experiment are shown.*

		K		
		K ₁	K ₂	
L	L ₁	15	15	30
	L ₂	15	15	30
		30	30	60

L—something that we could not do for the two two-groups experiments. Although we may look at the information about the interaction as pure “gravy,” we should note that some hypotheses may be constructed specifically to test for interactions. Thus, it may be that the experimenter is primarily interested in the interaction, in which case the other information may be regarded as “gravy.” But whatever the case, it is obvious that the factorial design yields considerably more information than separate two-groups designs and at considerably less cost to the experimenter. Still other advantages of the factorial design could be elaborated, but they might well come in your more advanced courses.

**SUMMARY OF AN ANALYSIS OF
VARIANCE AND THE COMPUTATION OF
AN F-TEST FOR A 2×2
FACTORIAL DESIGN**

Assume that the following dependent variable scores have been obtained for the four groups in a 2×2 factorial design.

		Condition A	
		A ₁	A ₂
Condition B:	B ₁	2	3
		3	4
		4	5
		4	7
		5	9
		6	10
		7	13
	B ₂	5	4
		6	6
		7	7
		8	9
		8	10
		8	11
		8	14

1. The first step is to compute ΣX , ΣX^2 , and n for each condition. The values have been computed for our example:

		Condition A	
		A ₁	A ₂
Condition B:	B ₁	$\Sigma X = 31$ $\Sigma X^2 = 155$ $n = 7$	$\Sigma X = 51$ $\Sigma X^2 = 449$ $n = 7$
	B ₂	$\Sigma X = 50$ $\Sigma X^2 = 366$ $n = 7$	$\Sigma X = 61$ $\Sigma X^2 = 599$ $n = 7$

2. Using Equation (10.1), we next compute the total SS :

$$\begin{aligned}
 \text{Total } SS &= (\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2) - \frac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4)^2}{N} \\
 &= (155 + 449 + 366 + 599) - \frac{(31 + 51 + 50 + 61)^2}{28} \\
 &= 238.68
 \end{aligned}$$

3. The over-all among SS is computed by substituting in Equation (10.2):

Between SS

$$\begin{aligned}
 &= \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \frac{(\Sigma X_4)^2}{n_4} - \frac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4)^2}{N} \\
 &= \frac{(31)^2}{7} + \frac{(51)^2}{7} + \frac{(50)^2}{7} + \frac{(61)^2}{7} - \frac{(31 + 51 + 50 + 61)^2}{28} \\
 &= 67.25
 \end{aligned}$$

4. The within SS is determined by subtraction, Equation (10.3):

$$\begin{aligned}
 \text{Total } SS - \text{over-all among } SS &= \text{within } SS \\
 238.68 - 67.25 &= 171.43
 \end{aligned}$$

5. We now seek to analyze the over-all among SS into its components, viz., the between A SS , the between B SS , and the $A \times B$ SS . The between A SS may be computed with the use of Equation (10.4).

Between A SS

$$\begin{aligned}
 &= \frac{(\Sigma X_1 + \Sigma X_3)^2}{n_2 + n_3} + \frac{(\Sigma X_2 + \Sigma X_4)^2}{n_2 + n_4} - \frac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4)^2}{N} \\
 &= \frac{(31 + 50)^2}{7 + 7} + \frac{(51 + 61)^2}{7 + 7} - \frac{(31 + 51 + 50 + 61)^2}{28} = 34.32
 \end{aligned}$$

The between B SS may be computed with the use of Equation (10.5).

Between B SS

$$= \frac{(\sum X_1 + \sum X_2)^2}{n_1 + n_2} + \frac{(\sum X_3 + \sum X_4)^2}{n_3 + n_4} - \frac{(\sum X_1 + \sum X_2 + \sum X_3 + \sum X_4)^2}{N}$$

Between B SS

$$= \frac{(31 + 51)^2}{7 + 7} + \frac{(50 + 61)^2}{7 + 7} - \frac{(31 + 51 + 50 + 61)^2}{28} = 30.04$$

The sum of squares for the interaction component ($A \times B$) may be computed by subtraction:

$$A \times B \text{ } SS = \text{over-all among } SS - \text{between } A \text{ } SS - \text{between } B \text{ } SS \\ 67.25 - 34.32 - 30.04 = 2.89$$

6. Compute the several degrees of freedom. In particular, determine df for the total source of variance Equation (10.7), for the overall among source Equation (10.8), and the within source Equation (10.9). Following this, allocate the over-all among degrees of freedom to the components of it; namely that between A , that between B , and that for $A \times B$.

$$\begin{aligned} \text{Total } df &= N - 1 \\ &= 28 - 1 = 27 \end{aligned}$$

$$\begin{aligned} \text{Over-all among } df &= r - 1 \\ &= 4 - 1 = 3 \end{aligned}$$

$$\begin{aligned} \text{Within } df &= N - r \\ &= 28 - 4 = 24 \end{aligned}$$

The components of the over-all among df are:

$$\begin{aligned} \text{Between } df &= r - 1 \\ \text{Between } A &= 2 - 1 = 1 \\ \text{Between } B &= 2 - 1 = 1 \end{aligned}$$

$$\begin{aligned} A \times B \text{ } df &= (\text{number of } df \text{ for between } A) \times (\text{number of } df \text{ for between } B) \\ &= 1 \times 1 = 1 \end{aligned}$$

7. Compute the various mean squares. This is accomplished by dividing the several sums of squares by the corresponding degrees of freedom. For our example these operations, as well as the results of the preceding ones, are summarized:

Source of Variation	Sum of Squares	df	Mean Square	F
Between A	34.32	1	34.32	4.81
Between B	30.04	1	30.04	4.21
$A \times B$	2.89	1	2.89	.40
Within groups	171.43	24	7.14	
Total	238.68	27		

Compute an F for each "between" source of variation. In a 2×2 factorial design there are three F -tests to run. The F is computed by dividing a given mean square by the within groups mean square (assuming the case of fixed variables). These F 's have been computed and entered in the above table.

9. Enter Table 9.11 to determine the probability associated with each F . To do this find the column for the number of degrees of freedom associated with the numerator and the row for the number of degrees of freedom associated with the denominator. In our example they are 1 and 24 respectively. The F of 4.81 for between A would thus be significant beyond the 5 per cent level, and accordingly we would reject the null hypothesis for this condition. The F between B (4.21) and that for the interaction (0.40), however, are not significant at the 5 per cent level; hence we would fail to reject the null hypotheses for these two sources of variation.

PROBLEMS

1. An experimenter wants to evaluate the effect of a new drug on "curing" psychotic tendencies. He investigates two independent variables, the amount of the drug administered and the type of psychotic condition. He decides to vary the amount of drug administered in two ways, none and 2 cc. The type of psychotic condition is also varied in two ways, schizophrenic and manic-depressive. Diagram the factorial design that he used.

2. In the above experiment the psychologist used a measure of normality as his dependent variable. This measure varies between zero and ten, where ten is very normal and zero is very abnormal. Seven subjects were assigned to each cell. The resulting scores for the four groups were as follows. Conduct the appropriate statistical analysis and reach a conclusion about the effect of each variable and the interaction.

PSYCHOTIC CONDITION			
<i>Schizophrenics</i>		<i>Manic Depressives</i>	
	<i>Did Not</i>		<i>Did Not</i>
<i>Received Drug</i>	<i>Receive Drug</i>	<i>Received Drug</i>	<i>Receive Drug</i>
6	2	5	1
6	3	6	1
6	3	6	2
7	4	7	3
8	4	8	4
8	5	8	5
9	6	9	6

3. How would the above design be diagrammed if the experimenter had varied the amount of drug in three ways (zero amount, 2 cc, and 4 cc), and the type of psychotic tendency in three ways (schizophrenic, manic-depressive, and paranoid)?

4. How would you diagram the above design if the experimenter had varied the amount of drug in four ways (zero, 2 cc, 4 cc, and 6 cc) and the type of subject in four ways (normal, schizophrenic, manic-depressive, and paranoid)?

5. A cigarette company is interested in the effect of several conditions of smoking on steadiness. They manufacture two brands, Old Zincs and Counts. Furthermore they make each brand with and without a filter. A psychologist conducts an experiment in which he studies two independent variables. The first is brand, which is varied in two ways (Old Zincs and Counts), and the second is filter, which is also varied in two ways (with a filter and without a filter). He uses a standard steadiness test as his dependent variable. Diagram the resulting factorial design.

6. In the above experiment the higher the dependent variable score, the greater the steadiness. Assume that the results came out as follows (ten subjects per cell). What conclusions did the experimenter reach?

OLD ZINGS		COUNTS	
<i>With Filter</i>	<i>Without Filter</i>	<i>With Filter</i>	<i>Without Filter</i>
7	2	2	7
7	2	3	7
8	3	3	7
8	3	3	8
9	3	3	9
9	4	4	9
10	4	4	10
10	5	5	10
11	5	5	11
11	5	6	11

7. An experiment is conducted to investigate the effect of opium and marijuana on hallucinatory activity. Both of these independent variables were varied in two ways. Seven subjects were assigned to cells and amount of hallucinatory activity was scaled so that a high number indicates considerable hallucination. Assuming that adequate controls have been realized, and that a 5 per cent level of significance was set, what conclusions can be reached?

SMOKED OPIUM		DID NOT SMOKE OPIUM	
<i>Smoked Marijuana</i>	<i>Did Not Smoke Marijuana</i>	<i>Smoked Marijuana</i>	<i>Did Not Smoke Marijuana</i>
7	5	6	3
7	5	5	2
7	4	5	2
6	4	4	1
6	3	4	1
5	3	4	0
4	3	3	0

APPENDIX

THE RATIONALE FOR SELECTING
ERROR TERMS

In a 2×2 factorial experiment we partition the total sum of squares into the sum of four component sums of squares, known as factor A, factor B, interaction AB, and within groups sum of squares. Tests of the null hypotheses are made in terms of the four mean squares obtained from these sum of squares. With the exception of the within groups mean square, each of the other mean squares may contain more than one source of variation (or component of variance). The expected value of each mean square is shown in Table 10.21. Let us consider the fixed model first, and so we shall read down the column of Table 10.21 for the expected mean square for that model. Note, for instance, that the expected mean square for independent variable A has two possible components of variance: that for the within variance (σ_W^2) and $2n$ times the measure of variation due to variable A (σ_A^2). For the present discussion we shall ignore the constants, like $2n$, and focus only on the variances. Reading further down the column, note that the two possible components for independent variable B are the within variance (σ_W^2) and a measure of variation due to variable B (σ_B^2). The interaction source of variation likewise may be due to the within variance and that for the interaction (σ_{AB}^2). Finally, we note the within variance as the only component for that source of variation. In running our F -tests let us say that we start by testing the $A \times B$ interaction. If that F is significant, we conclude that there is an interaction between variables A and B; otherwise, there is no such interaction. Now we can more clearly see what happens when we conduct this F -test. That is, from our sample values we compute a mean square for the among source of variation for interaction, and likewise for the within groups. These mean squares are, roughly, estimates of the (population) variances for the interaction and for within groups. To compute the F ratio for interaction we divide the mean square for within groups into the mean square for interaction. Now, if the null hypothesis is true, $\sigma_{AB}^2 = 0$. Consequently, the interaction component of the mean square will be zero in the long run, so that the F ratio for interaction will consist only of an estimate for the within groups variance in the numerator, and also an estimate for the within groups variance in the denominator, yielding a value of approximately one. But if the null hypothesis is false, there is a positive value for the interaction variance. In this case the numerator of the F ratio should contain not only a within groups mean square but also a contribution for the interaction; hence F should be greater than one — it should be significant, in fact.

Table 10.21. *Expected Mean Squares for the Fixed, Random, and Mixed Models for a 2 × 2 Factorial Design.*

SOURCE OF VARIATION	EXPECTED MEAN SQUARES FOR THE			
	Fixed Model	Random Model	Mixed Model (A Fixed)	Mixed Model (B Fixed)
Independent Variable A:	$\sigma_W^2 + 2n\sigma_A^2$	$\sigma_W^2 + n\sigma_{AB}^2 + 2n\sigma_A^2$	$\sigma_W^2 + n\sigma_{AB}^2 + 2n\sigma_A^2$	$\sigma_W^2 + 2n\sigma_A^2$
Independent Variable B:	$\sigma_W^2 + 2n\sigma_B^2$	$\sigma_W^2 + n\sigma_{AB}^2 + 2n\sigma_B^2$	$\sigma_W^2 + 2n\sigma_B^2$	$\sigma_W^2 + n\sigma_{AB}^2 + 2n\sigma_B^2$
A × B Interaction:	$\sigma_W^2 + n\sigma_{AB}^2$	$\sigma_W^2 + n\sigma_{AB}^2$	$\sigma_W^2 + n\sigma_{AB}^2$	$\sigma_W^2 + n\sigma_{AB}^2$
Within Groups:	σ_W^2	σ_W^2	σ_W^2	σ_W^2

In short, when the (true) variance interaction is zero, we expect an F value of one. But if there is a component for interaction, F becomes greater than one. Furthermore, if the ratio $\frac{\text{Interaction mean square}}{\text{Within mean square}}$ is sufficiently great we say that F is significant. In summary, we can see why the within mean square is the appropriate error term for testing the interaction mean square for a fixed model: *Since the interaction mean square may contain the within variance and the interaction variance, we divide that mean square by the within mean square. If the resulting value is significantly greater than one, we conclude that there is a significant interaction.*

To continue to the source of variation for independent variable B we can see that precisely the same reasoning applies. Since this expected mean square contains the within variance as well as the variance due to B, we divide this mean square by the within mean square. The case for independent variable A is precisely the same. This reasoning will become clearer as we take up the random model.

For the random model, let us read down the column for the expected mean square for the case of random variables. As before, we note that the within source of variation has only one component, viz., the within groups variance (σ_W^2). Similarly, the $A \times B$ interaction is the same for the random as for the fixed model. We therefore test the interaction for the random model as before. We divide that mean square by the within groups mean square. The difference between the two cases occurs when we note that there are *three* variance components for independent variable B: For the random model it may contain not only the within groups variance (σ_W^2) and ($2n$ times) the variance due to B (σ_B^2), but also (n times) the interaction variance. Hence, we need to "divide out" both the within and the interaction variances for this expected mean square. Since these are the two variance components for the $A \times B$ interaction, we use the mean square for interaction as our error term for the independent variables. As before, if the variance for independent variable B is zero, as assumed by the null hypothesis, the numerator of the F ratio does not (in the long run) contain a mean square for factor B, and the expected value of F approximates one. But this F value increases as σ_B^2 increases; if the sample value is sufficiently great, the F is significant and we can reject the null hypothesis. The same reasoning applies for variable A.

We can, incidentally, now note the error that would occur should the incorrect error term be used. For example, in the random model, if you happen to divide the mean square for independent variable A by the within groups mean square, your F value will be artificially inflated. That is, the numerator of the F -test would contain a variance component for the interaction *and* a variance component for variable B. Your interest is in whether or not variable B is effective, but what you actually would be testing is the significance of the variance for interaction plus the variance for variable B. In this

instance if your F is significant, you do not know whether it is because the interaction or variable B is significant, or whether it is a combination of these sources of variation. Experimenters have actually concluded that variable B was significant when they have used the wrong error term. You now will not make that mistake.

We now can hastily deal with the case of mixed variables, for our principles have already been established. First let us consider the case where independent variable A is fixed and independent variable B is random. In Table 10.21 we read down the column labeled "Mixed Model (A Fixed)." Here the proper error term for testing the interaction source of variation is the within mean square, because the interaction mean square may contain a within variance component and an interaction variance component. Similarly, the mean square for the random variable (variable B) may contain a component due to within and a component due to variable B; hence, to test B we use the within groups mean square in the denominator of the F ratio. But the mean square for fixed variable A may contain sources of variation due to A and due to the within and due to interaction; consequently to test A, one uses the interaction mean square in the denominator of the F ratio.

Finally, for the mixed model where B is fixed (the last column of Table 10.21), we can apply precisely the same reasoning as for the mixed model where A is fixed. In this instance, one tests the interaction mean square by dividing that value by the within mean square. Similarly, the random variable (variable A) is tested by using the within groups mean square as the error term, but the fixed variable (variable B) is tested by using the interaction mean square for the error term.

EXPERIMENTAL DESIGN

Within-Subjects Design

The two-randomized-groups design, the matched-groups design, the more-than-two randomized groups design, and the factorial design are all examples of “between-subjects designs.” This is so because two or more values of the independent variable are selected for study, and *one* value is administered to *each* group in the experiment. We then compute the mean dependent variable score for each group, compute the difference *between* groups, and thus assess the effect of varying the independent variable. An alternative to a between-subjects design is a “within-subjects design” in which two or more values of the independent variable are administered, in turn, to the same subject. A dependent variable score is then obtained for each subject’s performance under each value of the independent variable; comparisons of these dependent variable scores under the different experimental treatments then allow assessment of the effects of varying the independent variable.

In short, for between-subjects designs we compare dependent variable scores *between* groups who have been treated differently. In within-subjects designs, the same subjects are treated differently at different times, and we

compare their scores as a function of different experimental treatments. For example, suppose that we wished to ascertain the effect of LSD on perceptual accuracy. Use of a between-subjects design would dictate (essentially) that we administer LSD to an experimental group, but not to a control group. We would then compare the mean scores of the two groups on a test of perceptual accuracy to determine possible effects of the drug. But for a within-subjects design we would administer the test of perceptual accuracy to the same subjects: (1) when they were under the influence of the drug; and (2) when they were in a normal condition. If the mean scores of the same subjects change as they go from one condition to the other, we ascribe the change in behavior to LSD.

One excellent example of the use of a within-subjects design is the classical experiment on memory performed by Ebbinghaus (1913). This pioneer, it will be recalled, memorized several lists of nonsense syllables and then tested himself for recall at various times after learning was completed. He then calculated the percentage of each list that was forgotten after varying periods of time. For example, he found that he had forgotten about 47 per cent of one list 20 minutes after he had learned it, that 66 per cent of a second list was forgotten after one day, that after two days he had forgotten 72 per cent of a third list, and so forth. By thus taking repeated measures on himself Ebbinghaus was able to plot amount forgotten as a function of the passage of time since learning and obtained his famous forgetting curve.¹

A number of other classic experiments have also been conducted with the use of a within-subjects design. In addition to much work on memory, for example, there has been a myriad of studies in the area of psychophysics (cf., Underwood, 1966; Woodworth and Schlossberg, 1955). Weber's experimentation that yielded the data necessary to formulate his famous law is but one case in point; you no doubt studied Weber's Law in introductory psychology. Let us now turn, however, to several ways in which this type of design is used in contemporary psychology.

TWO CONDITIONS, MANY SUBJECTS

We already have some familiarity with the *t* and *A*-tests for matched groups, so this provides us with a good basis for studying one kind of within-subjects design. In this case a measure is obtained for each of a number of subjects when they perform a certain task or under some given experimental

¹It is fortunate, incidentally, that Ebbinghaus was not a professional psychologist. For if he had been, he would have known that what he accomplished was impossible—psychologists of his time, for example, held that the "higher mental processes" (e.g., memory) were not susceptible to experimental attack.

condition. The same measure is taken again when the subjects perform another task, or under another experimental condition. A mean difference between each pair of measures is computed and tested to determine whether it is significantly different from zero. If this difference is not significant then it can not be concluded that the variation of the independent variable resulted in behavioral changes. Otherwise, the conclusion is in the affirmative. For example, McGuigan, Ostrov, and Savukas (1967) sought to test a hypothesis based on the work of Lepley (1952). Lepley's findings suggested that individuals engage in covert oral behavior ("subvocal speech") when they write words. The experiment thus called for direct recording of the amplitude of speech muscle responding during handwriting; among the measures recorded was chin electromyograms (EMG). These experimenters had their subjects relax and then systematically either draw ovals or write words. It was reasoned that the motor task of drawing ovals, which does not involve the use of language, would serve as a control condition, i.e., subjects are generally active when they write, but is there a greater increase in covert oral behavior during writing than occurs during comparable activity that is not language in nature? To answer this question, the experimenters first determined the increase in chin EMG during writing, i.e., they subtracted amplitude of chin EMG during resting from that during writing for each subject. Then they similarly determined the increase in chin EMG amplitude for each subject during the drawing of ovals. The results are presented in Table 11.1. For example, Subject 1 increased her amplitude of covert oral responding (by this measure) during writing by $23.5 \mu\text{v}$ (μv = microvolts, which is one one-millionth of a volt); the comparable increase for this subject when she drew ovals was $12.0 \mu\text{v}$. And so on for the other subjects. The question is: Is there a significantly greater increase during the writing period than during the "ovals" period? To answer this question we compute the difference in response measures; for Subject 1 we can see that the difference is $11.5 \mu\text{v}$. To conduct a statistical test, we next compute the sum of the differences and the sum of the differences squared, values that are included in Table 11.1. If the mean of the difference scores is not significantly greater than zero, we will not be able to assert that variation of the experimental tasks produced a change in covert oral behavior. The appropriate test is either the matched t -test or the A -test; since the latter is easier to compute, we select it so that, hopefully, the saved time can be used to good advantage elsewhere. Recall from p. 178 that:

$$(11.1) \quad A = \frac{\sum D^2}{(\sum D)^2}$$

Substitution of the appropriate values from Table 11.1 results in:

$$A = \frac{28,578.34}{(384.00)^2} = \frac{28,578.34}{147,456.00} = .194$$

Table 11.1. *Change in Chin Electromyograms (μv) During Handwriting and While Drawing Ovals.*

Subject No.	Handwriting	Drawing Ovals	Difference
1	23.5	12.0	11.5
2	.3	5.8	-5.5
3	86.8	52.8	34.0
4	33.3	-29.3	62.6
5	46.4	22.9	23.5
6	-1.6	-24.1	22.5
7	26.2	-20.7	46.9
8	6.6	6.0	.6
9	16.9	-13.1	30.0
10	43.6	22.6	21.0
11	143.6	6.7	136.9

$\Sigma D = 384.0$
 $\Sigma D^2 = 28,578.34$

Referring to Table 8.12, on p. 177, we find that this A (with 10 df) indicates that the mean of the differences between the two conditions is significantly different from zero, i.e., $P < .05$ (P actually would have been less than .02 had we set this as our level of significance). The conclusion, thus, is that the subjects emitted a significantly larger amplitude of covert oral responding during handwriting than during a comparable motor task that was non-language in nature (than drawing ovals). The interpretation of this finding is that individuals engage in covert *language* behavior when receiving and processing language stimuli (words).

Incidentally, the question on which we focused was: Is there a greater change in the dependent variable when the subjects engaged in task A than in task B? Often, as in this case, performance in the two tasks is ascertained by comparison with some standard condition, such as during a resting state. In this event another, but related, question can also be asked, viz., did performance under condition A (and B) change significantly from the standard condition? The data in Table 11.1 can also provide answers to these questions. Since the scores under "Handwriting" and "Drawing Ovals" are themselves different scores, they can also be analyzed by the A -test. That is, you will recall that a measure was obtained for each subject when she was resting, and then when she was writing. The score 23.5 for Subject 1 thus was obtained by subtracting the resting level from the level during writing. To determine whether there was a significant increase in covert oral behavior when the subjects changed from resting to writing, one merely needs to compute the sum of the scores under the "handwriting" column, and the sum of the squares of these scores. Then substitute these values into Equation (11.1) and ascertain whether the A value is significant. Is it? How about the scores for the "ovals" condition?

SEVERAL CONDITIONS, MANY SUBJECTS

The within-subjects design in which two experimental treatments are administered to the same subjects can be extended indefinitely. Let us briefly illustrate one extension by considering an experiment by Underwood (1945) in which four values of the independent variable were administered to the same group of subjects. First, all subjects were systematically presented with the following tasks: (1) they studied no lists; (2) they studied (for four trials) two lists of paired adjectives; (3) they studied four lists of paired adjectives; and (4) they studied six such lists. Following this they completely learned another list of paired adjectives; 25 minutes later they were tested on this list, and the dependent variable was the number of paired adjectives that they could correctly recall. The results are presented in Fig. 11.1, where

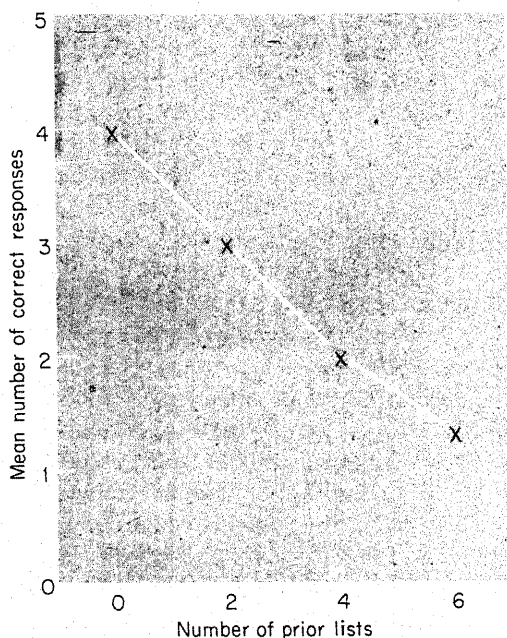


FIGURE 11.1.

The larger the number of lists studied before learning, the greater the amount of proactive inhibition. (After Underwood, 1945).

it can be noted that the fewer the number of prior lists studied, the better the recall. As you no doubt have observed, this was an experiment on proactive inhibition (interference), i.e., when subjects study something then learn some other (related) material, the first learned material inhibits the recall of

the later learned material. Put another way, earlier learned material proactively interferes with the retention of later learned material, and Underwood showed that the greater the number of prior lists learned, the greater is the amount of proactive inhibition. But regardless of the subject matter findings, the point here is that subjects can be administered a number of experimental treatments by means of the within-subjects design.

EVALUATION OF WITHIN-SUBJECTS DESIGNS

The contrast has been between within- and between-subjects designs, and since we have presented both kinds it must be apparent that there are pluses and minuses for each. Let us take up some of the more straightforward points first:

1. First, it should be clear that the within-subjects design has a great advantage as far as economy of subjects is concerned. For in this design scores are available for all subjects under all treatment conditions. The simplest illustration is the contrast with a two-randomized-groups design; you have two values of the independent variable and, say, 20 subjects in each group for a total of 40 subjects. Hence, for each condition, you have 20 scores. But in a within groups design you would have all subjects serve under both conditions; therefore you could either: (1) run a total of only 20 subjects to obtain the number of scores equal to that for the between groups design; or (2) you could still run 40 subjects and have twice as many data for each treatment condition.

2. The within-groups design is also relatively advantageous if your experimental procedure demands a lot of time or energy in getting ready to run your subjects. For example, if you are conducting psychophysiological research, it takes a fair amount of time and patience to properly attach the necessary electrodes, etc. Or, suppose that you are conducting neuropsychological research, such as implanting brain electrodes in animals. Once you made such a sizeable investment in your preparation, you probably would want to collect a lot of data, probably by running your subjects under a variety of conditions.

3. The advantage of the within-groups design that is most frequently cited is that, by its means, you typically reduce your error variance. We have seen in Chapter 8 that matching subjects on an initial measure can result in a sizeable increase in the precision of your experiment. The same logic applies here. In effect, by taking two measures on the same subject, you can reduce your error variance in proportion to the extent to which the two measures are correlated. Put another way, one reason that the error variance may be

large in a between-groups design is that it includes the extent to which individuals differ. But since, in a within-subjects design, you repeat your measures on the same subjects, you eliminate individual differences from your error variance. Hence, rather than having an independent control group, you have each subject serve as his own control.

You are probably getting suspicious by now, wondering "what's the rub?" There are some:

4. This is a relatively minor one. Sometimes one of the two (or more) treatments is such that if it comes first, it is not reasonable to then present the other. Suppose that, as in the Jacobsen et al. experiment on p. 226, you wish to study the effects of injecting RNA into an organism. Further, suppose that you wish to use as a control group, animals that did not receive RNA. Clearly a between-subjects design has to be used here, for you couldn't first administer RNA, test the animals, and then take RNA out of them and retest them. This brings us face to face with a topic that has been lurking in the background throughout this chapter, viz., the problem of the order in which the experimental treatments are presented to the same subjects.

5. This is the major problem entailed by the use of a within-subjects design. The one technique that we have discussed for systematically presenting the order of conditions is counterbalancing. If you do not adequately recall this discussion, you should refresh yourself now (p. 132-135).

You have decided, let us say, to use a within-subjects design and now face the question of what order to present your treatments to your subjects. If you know that the order of conditions will have no effect on your dependent variable, and that there are no practice and fatigue effects, then you have no problem — whether or not you use a counterbalanced design is irrelevant. Assuming that you are in this fortunate position, you clearly should use a within-groups design. This is, however, more or less of a "thank you for nothing" answer, for unless you have a massive amount of data on your particular variables, you would never know that you are in this happy state. So we must be more realistic.

First, let us be clear about the seriousness of the problem. If you assume that one condition does not interact with another, when in fact it does, your conclusions can be drastically distorted. An excellent example of this error is pointed out by Underwood (1957a) in which Ebbinghaus' forgetting curve is reexamined. Let us emphasize here that Ebbinghaus learned a number of lists that were to be recalled at later times. Implicit in Ebbinghaus' assumptions, as we look back from our present vantage point, was that his treatments did not interact to affect his dependent variable. Put more simply, the assumption was that the learning of one list of nonsense syllables did not affect the recall of another. The results of Ebbinghaus' research indicated that the large majority of what we learn is very rapidly forgotten, e.g., after one day, according to Ebbinghaus' forgetting curve, about 66 per cent is

forgotten. The consequence of this research, incidentally, has been sizeable and has been the source of great discouragement to educators for many decades. However, we now know that the basic assumption of Ebbinghaus' experimental design is not tenable. In fact, one of the major advances in the study of memory has been to establish the great effect of competition amongst materials that have been learned; the result has been the interference theory of forgetting. It was Underwood (1957a) who astutely demonstrated the defect in Ebbinghaus' design. For he showed that Ebbinghaus, by learning a large number of lists, created a condition in which he maximized amount of forgetting. When we take the number of previous lists learned into account, the situation is quite different. Figure 11.2 vividly makes the point, for this

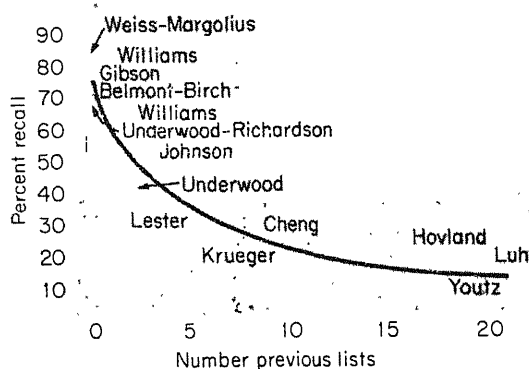


FIGURE 11.2.

Recall as a function of number of previous lists learned. (After Underwood, 1957a).

forgetting curve, plotted by Underwood, indicates the per cent forgotten after 24 hours as a function of number of previous lists learned. We can note that the situation is really not as bad as Ebbinghaus' results would have us believe. True, when a number of lists are learned, forgetting is great. But if there have been no previous lists learned, only about 25 per cent is forgotten after one day. The lesson thus should be clear: By using a within-groups design Ebbinghaus gave us a highly restricted set of results that were greatly overgeneralized and that thus led to erroneous conclusions about forgetting. Had he used a between-subjects design in which the subject learned only one list, he would have concluded that the amount forgotten was relatively small.

Let us briefly consider one more illustration of the difference in results that can be obtained by using the two types of designs. Grice and Hunter (1964) varied the intensity of the conditioned stimulus in a conditioning experiment, but they used two designs. In the between-subjects design one

group received a low intensity conditioned stimulus (soft tone), while a second group received a high intensity conditioned stimulus (loud tone). They also conducted the experiment using a within-subjects design in which all subjects received both values of the conditioned stimulus. The question was: Did variation of the intensity of the conditioned stimulus affect the strength of the conditioned response? The results are presented in Figure 11.3.

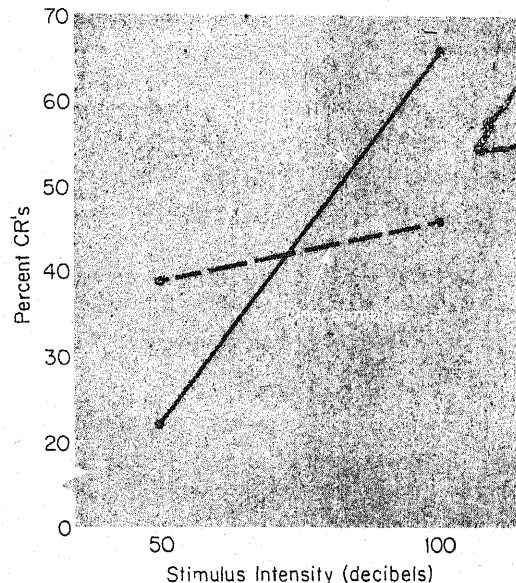


FIGURE 11.3.

Per cent CRs during the last 60 trials to the loud and soft tones under the one- and two-stimulus conditions.

We can see that in both experiments, there was an increase in the percentage of conditioned responses made to the conditioned stimulus as the intensity of the conditioned stimulus increased from 50 to 100 decibels (db). But the slopes of the curves are dramatically different. The difference in per cent of conditioned responses as a function of stimulus intensity was not significant for the between-groups design ("one stimulus"), while it was for the within-subjects design (in which the subjects received "two stimuli"). In fact, the magnitude of the intensity effect is more than five times as great for the two-stimuli condition than for the one-stimulus condition. Hence, the dependent variable scores were influenced by the number of conditions in which the subjects served; there was an interaction between stimulus intensity and number of presentations of stimuli.

With this appreciation of the importance of the possible interaction effects of our treatments, let us now return to the question of the order to use in a

within-subjects design. The purpose of counterbalancing, we have said, is to control practice and fatigue effects — to distribute them equally to all experimental conditions. But, we pointed out, by thus controlling these variables, you might inherit problems of a different sort, viz., asymmetrical transfer effects. Hence, if you use a counterbalanced design, you should demonstrate (by appropriate statistical analysis) that there was no differential transfer among your conditions. On the other hand, if you expect (fear, might be a better word) asymmetrical transfer effects, you should seek some other method for presenting your treatments. In an excellent article by Gaito (1961), six cases for presenting repeated treatments to the same subjects are examined. There are potentially serious methodological problems with five of the designs, he points out, and these should be used with caution; several methods of counterbalancing are among these five designs, as you no doubt anticipated. The sixth design is methodologically sound; it calls for the randomization of the order of the treatments. For example, if you have three treatments (A, B, and C), and all subjects are to receive all treatments, then you randomly determine the order of A, B, and C for each subject. However, Gaito points out, the error variance in the design using randomization of treatments "may be quite large."

In summary, it is quite clear that there are several important advantages of the within-subjects design over the between-subjects design. But if you cannot effectively handle the control problems entailed by counterbalancing, then you have two alternatives: (1) present your treatments to your subjects in a random order; or (2) use a between-subjects design. If you select the latter, it is interesting to note that Gaito suggests serious consideration of the matched-groups design (Chapter 8).

A SINGLE-SUBJECT DESIGN WITH REPLICATION

The designs on which we have thus far concentrated have, in general, been based on the assumption that many subjects would be studied for a short period of time. The effects of varying the independent variable were studied by computing group means and evaluating them relative to the amount of error in the experiment. If changes in group means were sufficiently larger than experimental error, it was concluded that there was a relationship between the independent and the dependent variables. For example, in designs analyzed by analysis of variance the value of the numerator of the F ratio is an indication of the effects of varying the independent variable, (the among-mean square) while the denominator (the within-mean square) is the error variance. The F -test yields a significant value if the numerator is

sufficiently larger than the denominator. In short, the strategy has been to attempt to determine whether changes in behavior produced by the independent variable were sufficiently great to show through the "noise" in the experiment. This strategy has been vigorously criticized by Skinner (e.g., 1959) and his associates. Although we shall take up the topic of error variance in greater detail in Chapter Fifteen, this brief introduction serves to explain the rationale of the present design. Skinner's methodology, which is referred to as "The Experimental Analysis of Behavior," is an effort to sizeably reduce the error variance in the experiment. From the present point of view, experimental error has two major components: (1) that due to individual differences among the subjects; and (2) that due to ineffective control procedures. Briefly, the former is eliminated in this design, simply put, by using only one subject in each experiment; the latter is reduced by establishing highly controlled conditions in the experimental situation. Rather than studying a number of subjects for a short period of time, Skinner has proposed that we study a single subject over an extended period.

The subject is placed in a well controlled environment, typically in a Skinner Box if it is an animal subject. In the Skinner Box an effort is made to eliminate or hold constant all extraneous stimuli so that it is sound deadened, lighting is constant, no unique olfactory cues are present, and so forth. Then the animal performs for a lengthy period during which time his baseline level of performance is established. For example, a hungry rat might be conditioned to press a lever and receive food. Conditioning procedures are continued until the rat displays a stable rate of bar pressing, as shown by his cumulative record.

The cumulative record is established as follows. The writing pen on an ink recorder is automatically activated each time the rat presses the bar. The pen writes on a continuously moving piece of paper and is elevated one unit for each bar press. Figure 11.4 shows this process. Imagine that the paper is moving from right to left. Then each bar press moves the pen up one unit. When no response is made, the pen indicates this by continuing to move horizontally. Hence, we can note that after one minute the rat made a response, that he did not make another response until two minutes had elapsed, that a third response was made after two and one half minutes, and so forth. If we wish to know the total number of responses made after any given time in the experimental situation, we merely read up to the curve from that point in time and over to the vertical axis. For example, we can see that after five minutes the rat had made five responses, as read off of the vertical axis. Incidentally, the cumulative response curve is a summation of the total number of responses made since time zero; this means that the curve can never decrease, i.e., after the rat has made a response, as indicated by an upward mark, that response can never be unmade; the pen can never move down. Think about this point, if the cumulative response curve is new to you.

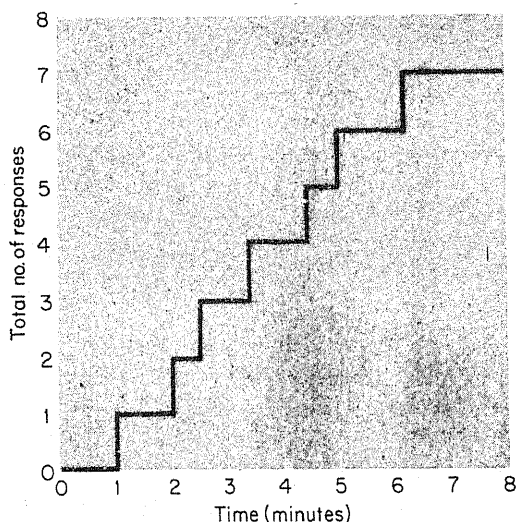


FIGURE 11.4.

A cumulative response curve shown in detail.

We have shown in Figure 11.4 a short portion of the cumulative response curve. The experimenter allows the subject to continue responding so that a considerable amount of his history is recorded. The animal eventually performs in a very stable fashion so that his response rate is quite constant. Now, once this steady state has been established, it is reasonable to extrapolate the curve indefinitely, so long as the conditions remain unchanged. It is at this time that the experimenter introduces some unique treatment. The logic is quite straightforward — if the response curve changes, that change can be ascribed to the effects of the new stimulus condition. Once it has been established that the curve changes, the experimental condition can be removed and, providing there are no lasting effects, the curve should return to its previously stable rate. Additional conditions can then be presented, as the experimenter wishes. It can, thus, be seen that this is a within-subjects design in which repeated treatments are administered to the same subject.

One of the interesting aspects of Skinner's research has been his concern for technological matters. Although his laboratory work yielded a number of valuable principles of behavior, he and his followers have made a strong effort to explore applications of the principles to the problems of everyday life. Let us illustrate the Skinnerian methodology more completely by considering a portion of an applied study conducted by Harris, Wolf, and Baer (1964). These experimenters studied a four year old boy who cried a great deal after he experienced minor frustrations. In fact, it was determined that he cried about eight times during each school morning. The cumulative

number of crying episodes can be studied for the first ten days of the experiment in Figure 11.5. The question was: What is the reinforcing event that maintains this crying behavior? The experimenters hypothesized that it was the special attention from the teacher that the crying brought. The paradigm is thus the same as that for the rat in the Skinner Box: When the response is made (the bar is pressed or the child cries), reinforcement occurs (food is delivered or the teacher comes to the child). After ten days when the response rate was stabilized, the experimental treatment was effected: For the next ten days the teacher ignored the child's crying episodes, but she did reinforce more constructive responses (verbal and self-help behaviors) by giving the child approving attention. As can be seen in Figure 11.5, the number of crying episodes sharply decreased with the withdrawal of the teacher's reinforcement for crying and during the last five of these ten days only one crying response was recorded. During the next ten days, reinforcement was reinstated — whenever the child cried, the teacher attended to the boy as she had originally done. Approximately the original rate of responding was reinstituted. Then, for the last ten days of the experiment, reinforcement was again withdrawn and the response rate returned to a near zero level, as you can note by the last series of light circles.² Furthermore, it remained there after the experiment was terminated. The experiment was repeated with another four year old boy, with the same general results.

This last point is very important in Skinner's methodology. Once it has been determined in an experiment with a single subject that some given treatment affects rate of responding, the experiment is replicated. When, under highly controlled conditions, it is ascertained that several other subjects behave in the same way to the change in stimulus conditions, the results are generalized to the population of organisms sampled. The point we made several times earlier in the book applies here too, i.e., the extent to which the results can be generalized to the population of organisms depends on the extent to which that population has been sampled. Unfortunately, researchers who use the present methodology often do not include a large enough sample of organisms in their experiment.

The methods used in the experimental analysis of behavior have much to recommend them. Skinner's work, and that inspired by him, has had a major influence on contemporary psychology. In addition to his contributions to pure science, it is important to emphasize that the results of using this methodology have had a sizeable impact in such technological areas as

²Skinner holds that statistics are not necessary with his methodology, a position that has led to some controversy. The changes in behavior, he holds, are sufficiently clear cut that statistical tests are not required to demonstrate an effect. This is fortunate if the changes in behavior *are* that apparent. However, one could easily conduct a statistical test to remove any doubt that an apparent change in behavior due to the introduction of an experimental treatment is, in fact, significant.

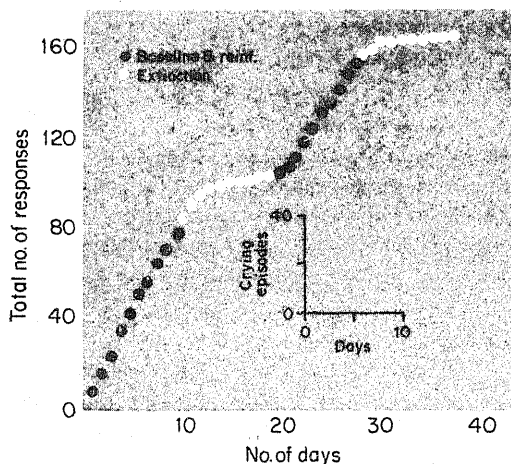


FIGURE 11.5.

Cumulative record of the daily number of crying episodes. The teacher reinforced crying during the first ten days (dark circles) and withdrew reinforcement during the second ten days (light circles). Reinforcement was reinstituted during the third period of ten days (dark circles) and withdrawn again during the last ten days. (After Harris, *et al.*, 1964.)

education (e.g., programmed learning), social control, psychotherapy, and so forth. It is likely that, should you continue to progress in psychology, you will find that you can make good use of this type of design. Particularly at this stage in our development we should encourage a variety of approaches in psychology, for we have many questions that appear difficult to answer. No single methodological approach can seriously claim that it will be universally successful, and we should maintain as large an arsenal as possible. Sometimes a given problem can be most effectively attacked by one kind of design, while another is more likely to yield to a different design.

THE LOGICAL BASES OF EXPERIMENTAL INFERENCES

We have said that an experimenter should state his hypothesis explicitly. His experiment has as its purpose the gathering of data that are relevant to the hypothesis. These data are summarized in the form of an evidence report, and the experimenter confronts the hypothesis with the evidence report. If the two are in agreement, he concludes that the hypothesis is confirmed. Otherwise it is not. Let us analyze the relationship between the hypothesis and the evidence report.

The hypothesis, preferably, is stated in the conditional form; that is, as an if-then relation. To pursue the example of Chapter 3, recall that the hypothesis that industrial work groups in great inner conflict have low production levels was stated as follows: *If* an industrial work group is in great inner conflict, *then* that work group will have a low production level. To test this hypothesis we might form two groups of subjects, one in great conflict and the other quite harmonious. We would then collect data on the production output of the two groups. Assume that the statistical analysis of the data shows that the in-conflict-group has a significantly lower production level than the harmonious group.

FORMING THE EVIDENCE REPORT¹

An evidence report is a summary statement of the results of an empirical investigation; it is a sentence that precisely summarizes what was found. In addition, the evidence report states that the antecedent conditions of the hypothesis were realized. Hence, the evidence report consists of two parts: a statement that the antecedent conditions of the hypothesis held, and that the consequent conditions were found to be either true or false. The general form for stating the evidence report is thus that of a conjunction. Recalling the general form of the hypothesis as "If a , then b ," a denotes the antecedent conditions of the hypothesis and b the consequent conditions. Hence, the possible evidence reports would be " a and b ," or " a and not b ," for the cases where the consequent conditions were found to be (probably) true and false respectively. The former is a positive evidence report and the latter is a negative one.

We shall illustrate by continuing the example. Let a stand for "an industrial work group is in great inner conflict" and b for "that work group has a lowered production level." In our hypothetical experiment we had an industrial work group that was in great inner conflict, and so we may assert that the antecedent conditions of our hypothesis were realized, that they were present in the experimental situation. Since our finding was that that work group had a lower production level than a control group, we may also assert that the consequent conditions were found to be true. Thus, our evidence report is: "An industrial work group was in great inner conflict *and* that work group had a lowered production level."

At this point, let us examine how we tell whether the consequent conditions of the hypothesis are true or false. In this example the group with inner conflict had a lower production level than did a control group. We needed the control group as a basis of comparison. For without such a basis "lower production level" does not mean anything — it must be lower than something. And so it is for all experiments. The way to tell whether or not consequent conditions are true is by comparing the results obtained under an experimental condition with those for some other condition, usually by means of a control group. That the hypothesis implicitly assumes the existence of a control or other group may be clarified by stating the hypothesis in the following manner: "If an industrial work group is in great inner conflict, then that work group will have a lower production level *than that of a group*

¹The term "evidence report" is taken from Hempel (1945) and is used because of its descriptive nature. Similar terms are "observational sentence," "protocol sentence," and "concept by inspection."

that is not in inner conflict. The direction of the comparison is determined by the consequent conditions of the hypothesis. In this example, the hypothesis states that the group with inner conflict should have a *lower* production level than a control group. If the statistical analysis indicates that it is significantly lower than the control group, we conclude that the consequent conditions are probably true. If the statistical analysis indicates that the group with the inner conflict has a production level that is significantly higher than the control group, however, or if there is no significant difference between the two levels, we conclude that the consequent conditions are probably false. The evidence report would then be: "An industrial work group was in great inner conflict and that work group did not have a lowered production level."

With this format for forming the evidence report before us, we shall now consider the nature of the inferences made from it to the hypothesis. Before we do this, however, it will be necessary to discuss the general topic of inferences and how they are made.

INDUCTIVE AND DEDUCTIVE INFERENCES

Let us say that we have a set of statements that we shall denote by *A*. These statements contain information on the basis of which we can reach another statement, *B*. Now, when we proceed from *A* to *B*, we make an *inference*. An inference, then, is a process by which we reach a conclusion on the basis of certain preceding statements — it is a process of reasoning whereby we start with *A* and arrive at *B*. We then have some belief that *B* is true on the basis of *A*. There are two kinds of inferences, inductive and deductive. In both types, our belief in the truth of *B* is based on the assumption that *A* is true. The essential difference between the two is that in an inductive inference, we reach the conclusion that, if *A* is true, *B* is true with some degree of probability; with a deductive inference, however, we can conclude that *B* is *necessarily* true if *A* is true.

The statement "Every morning that I have arisen, I have seen the sun rise" might be *A*. On the basis of this statement, we may infer the statement *B*: "The sun will always rise each morning." Now, does *B* necessarily follow from *A*? It *does not*, for while you may have always observed the rise of the sun in the past, it does not follow that it will *always* rise in the future. *B* is not *necessarily* true on the basis of *A*. Although it may seem unlikely to you now, it is entirely possible that one day, regardless of what you have observed in the past, the sun will not rise. We can only say then, that *B* has some degree of probability (is probably true) on the basis of the information contained in *A*. When we make an inference that may be in error, we say that the result of the inference (the conclusion) can only have a certain probability of being

true. Thus, the probability that statement B can be (inductively) inferred to be true on the basis of statement A may be high, medium, or low. The fact that we make inferences from one statement to another with a certain degree of probability sometimes leads us to use the term *probability inference* as a synonym for *inductive inference*.

It is possible (at least in principle) to determine the probability of an inference precisely as a specific number, rather than simply to say "high" or "low." Conventionally, the probability of an inductive inference may be expressed by any number from zero to one.² Thus, we may say that the probability (P) of the inference from A to B is 0.40, or 0.65, or whatever. Furthermore the closer P is to 1.0, the higher the probability that the inference will result in a true conclusion (again, assuming that A is true). By analogy, the closer P is to 0.0, the lower the probability that the inference will result in a true conclusion, or, if you will, the higher the probability that the inference will result in a false conclusion. At this point we may note that if the probability of an inference is 1.0, the conclusion is necessarily true if the statements on which it is based (i.e., A) are true; and if P is 0.0, the conclusion is necessarily false. As we have previously noted, however, neither of these situations obtain when an inductive inference is made.

We may thus say that if the probability that B follows from A is 0.99, it is rather sure that B is true. The previous example of the inference from "Every morning that I have arisen, I have seen the sun rise" to "the sun will always rise each morning" is an example of an inference with a high probability. The probability of this inference is, in fact, extremely close to, but still not quite, 1.0. On the other hand, if the probability of the inference from A to B is 0.03, we know that this probability is extremely low and thus are not likely to accept B as true on the basis of A . For example, the probability of the inference from the statement "a person has red hair" to the conclusion that "that person is very temperamental," would be one with a very low probability.

In short, then, an inductive inference is made when one passes from one statement to another with a lack of certainty; he infers that one statement probably implies another. And we may express our degree of belief in the truth of the results of such an inference by the use of a number that varies from 0.0 to 1.0. The higher the number, the greater the probability that the inference results in a true conclusion.

We said that a deductive inference is made when the truth of one statement is necessary, based on another one or set of statements, if statement A necessarily implies B . In this case the inference is strict. Consider the following

²To emphasize that 0 to 1 is an arbitrary range we may note that some authors allow P to assume any value between -1 and $+1$.

statements as an example of a deductive inference. We might know that "all anxious people bite their nails" and further that "John Jones is anxious." We may, therefore, deductively infer that "John Jones bites his nails." In this example, if the first two statements are true (they are called premises), the final statement (the conclusion) is necessarily true.

The determination of whether or not a given inference is deductive or inductive lies in the realm of logic. In both deductive and inductive (probability) logic certain rules (known as rules of inference or transformation) have been developed that indicate how to proceed from one set of statements to another. If an inference conforms to the rules of deductive logic, it is deductive. If it conforms to the rules of inductive logic, it is inductive. Since this is not a book in logic, we shall not explore the various rules to any great extent. We shall simply indicate the rules that are used in making inferences from evidence reports to various types of hypotheses. In doing this, we shall indicate which are valid inferences, and whether they are inductive or deductive.

DIRECT VS. INDIRECT STATEMENTS

The statements with which science deals may be divided into two categories, direct or indirect. A direct statement is one that refers to limited phenomena that are immediately observable, that is, phenomena that can all be observed directly with the senses. For example, the statement "that bird is red" is direct. Of course the use of various kinds of auxiliary apparatus (e.g., microscopes, telescopes) to aid the senses may be used in forming direct statements. Hence, the statements "there is a sun spot" or "there is an amoeba" are also direct statements, since a person's sensory apparatus can be extended by various types of equipment. The procedure for testing a direct statement is straight-forward: Compare the statement with an observation of the phenomenon with which it is concerned. More precisely, compare the statement with the evidence report. If the two are in agreement, the statement may be regarded as true; otherwise it is false. For example, in testing the direct statement "That door is open," we observe the door. If we find that it is open, our observation agrees with the direct statement, and we conclude that the statement is true. If we observe the door to be closed, we conclude that the direct statement is false.

An indirect statement is one that cannot be directly tested. Such statements usually deal with phenomena that cannot be directly observed (logical constructs, such as atoms, electricity, or habits) or that are so numerous or extended in time that it is impossible to view them all. A universal hypothesis is of this type — "All men are anxious." It is certainly impossible to observe all men (living, dead, and as yet unborn) to see if the statement is true. The

universal hypothesis is the type in which scientists are most interested, since it is an attempt to say something about one or several variables for all time, at all places.³

Clearly, then, indirect statements cannot be directly tested. In order to test indirect statements it is necessary to reduce them to direct statements. This reduction is accomplished by following the rules for deductive inferences. Consider an indirect statement S . By drawing deductive inferences from S we may arrive at certain logical consequences, which we shall denote s_1 , s_2 , and so forth. Now among the statements s_1 , s_2 , etc., we expect to find at least some that are direct in nature. Such statements may be tested by comparing them with appropriate evidence reports. Now, if these directly testable consequences of our indirect statement are found to be true, we may make a further (inductive) inference that the indirect statement itself is probably true. That is, while we cannot directly test an indirect statement, we can derive deductive inferences from such a statement and directly test them. If such directly testable statements turn out to be true, we may inductively infer that the indirect statement is probably true (see note 1 at the end of this chapter). But if its consequences turn out to be false, we must infer that the indirect statement is also false. In short, indirect statements that have true consequences are themselves probably true, but indirect statements that have false consequences are themselves false. This procedure is represented in Figure 12.1, although it will be necessary to analyze it more thoroughly later.

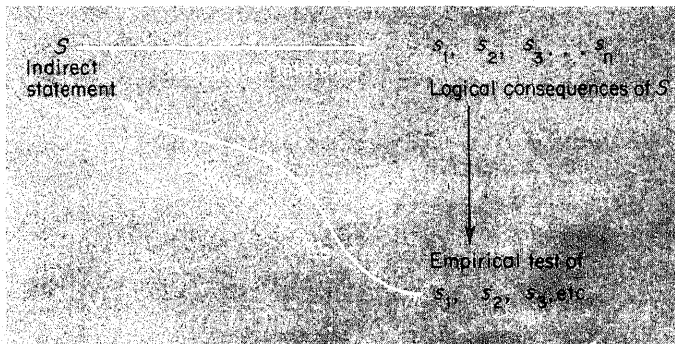


FIGURE 12.1.

Representation of the procedure for testing indirect statements.

To illustrate this procedure let us consider the universal hypothesis, "All men are anxious." Assume we know that "John Jones is a man," and "Harry

³Don't get too universal though, as one student did who defined a universal statement as a "relationship between *all* variables for all time and for all places."

Smith is a man." From these statements (premises) we can state (by deductive inference) that "John Jones is anxious" and "Harry Smith is anxious." Since the universal hypothesis is an indirect statement, it cannot be directly tested. However, the deductive inferences derived from this indirect statement are directly testable. We only need to determine the truth or falsity of these direct statements. If we perform suitable empirical operations and thereby conclude that the several direct statements are true, we may now conclude, by way of an *inductive inference*, that the indirect statement is confirmed.

The class of variables with which the indirect statement deals is of infinite number. For this reason it is impossible to test all the logically possible consequences of that indirect statement (e.g., we cannot test the hypothesis for all men). Further, it is impossible to make a deductive inference from the direct statements back to the indirect statement — rather, we must be satisfied with an inductive inference. And we know that an inductive inference is liable to error; its probability is less than 1.0. As long as we seek to test indirect statements, we must be satisfied with a probability estimate of their truth. We will never know absolutely that they are true.

CONFIRMATION VS. VERIFICATION

Our goal is to determine whether a given universal statement is true or false. To accomplish this goal we reason thusly: *If* the hypothesis is true, *then* the direct statements that are the result of deductive inferences are also true. Now, if we find that the evidence reports are in accord with the logical consequences (the direct statements), we conclude that the logical consequences are true. And if the logical consequences are true, we inductively infer that the hypothesis itself is probably true.

Note that we have been cautious and limited in our statements about concluding that a universal hypothesis is false. Under certain circumstances it is possible to conclude that a universal hypothesis is strictly false (not merely improbably or probably false) on the assumption that the evidence report is reliable. More generally (i.e., with regard to any type of hypothesis), it can be shown that under certain circumstances it is possible strictly to determine that a hypothesis is true or false, rather than probable or improbable, but always on the assumption that the evidence report is true. We will here distinguish between the processes of *verification* and *confirmation*. By verification we mean a process of attempting to determine that a hypothesis is strictly true or strictly false; confirmation is an attempt to determine whether a hypothesis is probable or improbable. This ties in with the distinction between inductive and deductive inferences. Under certain conditions it is possible to make a deductive inference from the consequence of a hypothesis (which has been determined to be true or false) back to that hypothesis.

Thus, where it is possible to make such a deductive inference, we are able to engage in the process of verification. Where we must be restricted to inductive inferences, the process of confirmation is used. To enlarge on this matter, let us now turn to a consideration of the ways in which the various types of hypotheses are tested (see note 2 in the Appendix).

INFERENCES FROM THE EVIDENCE REPORT TO THE HYPOTHESIS

Universal Hypotheses. Recall that the universal hypothesis specifies that all things referred to in the hypothesis have a certain characteristic. For such hypotheses we have indicated a preference for the "If a , then b " form. With such a form it is understood that we are referring to all a and all b . For example, if a stands for "rats are reinforced at the end of their maze runs" and b for "those rats will learn to run that maze with no errors," it is understood that we are talking about all rats and all mazes on which those rats might be trained. The general procedure for testing this hypothesis may be represented as follows (see note 3 in the Appendix):

Hypothesis:	If a then b	
Evidence Report:	a and b	
	↓	(Inductive Inference)
Conclusion:	"If a then b " is probably true.	

To illustrate this procedure by means of our example we might form two groups of rats; Group E is reinforced at the end of each maze run, but Group C is not. Let us say that at the end of a certain number of trials Group E is able to run the maze with no errors, but Group C is still making errors. A t -test indicates that Group E's performance is significantly superior to that of Group C. We are thus able to assert that the antecedent conditions of the hypothesis were realized and that the data were in accord with the consequent conditions. The evidence report is positive. The inferences involved in the test of this hypothesis may be illustrated as follows:

Universal Hypothesis:	If rats are reinforced at the end of their maze runs, then those rats will learn to run that maze with no errors.
Positive Evidence Report:	A (specific) group of rats was reinforced at the end of their maze runs and those rats learned to run the maze with no errors.
Conclusion:	The hypothesis is probably true.

We have attempted to show some specific steps in testing a hypothesis, to give you some insight into the various inferences that must be made for this

purpose. In your actual work, however, you need not specify each step, for that would become cumbersome. Rather, you should simply rely on the brief rules that we present for testing each type of hypothesis. To summarize the rule for testing a universal hypothesis for the case of a positive evidence report, we shall merely say that when the evidence report agrees with the hypothesis, that hypothesis shall be said to be confirmed.

To understand the test of the universal hypothesis when the evidence report is negative, we refer to the distinction between confirmation and verification. Clearly, the case of confronting the universal hypothesis with a positive evidence report is an example of confirmation. When the evidence report is negative, however, we are able to apply the procedure of verification. This is possible because the rules of deductive logic tell us that a deductive inference may be made from a negative evidence report to a universal hypothesis. The procedure for this situation may be illustrated as follows:

Universal Hypothesis:	If a then b	
Evidence Report:	a and not b	
	↓	(Deductive Inference)
Conclusion:	"If a then b " is false	

For example:

Universal Hypothesis:	If rats are reinforced at the end of their maze runs, then those rats will learn to run that maze with no errors.
Negative Evidence Report:	A group of rats was reinforced at the end of their maze runs and those rats did not learn to run that maze without any errors.
Conclusion:	The hypothesis is false (see note 4 in Appendix).

We may thus see that it is possible to determine that a universal hypothesis is (strictly) false (through verification) if the evidence report is negative. But if the evidence is positive, we cannot determine that the hypothesis is (strictly) true; rather, we can only say that it is probable (through confirmation). This characteristic of being able to verify a hypothesis in only one direction (for example being able to determine that it is strictly false, but not being able to determine that it is strictly true) has been called by Reichenbach (1949) *unilateral verifiability*. And we shall see that it is characteristic of all universal and existential hypotheses.

Existential Hypotheses. This type of hypothesis says that there is at least one thing that has a certain characteristic. Our example, stated as a positive existential hypothesis would be: "There is a (at least one) rat that, if it is reinforced at the end of its maze runs, then it will learn to run that maze with no errors." If this type of hypothesis is strictly true, then it can be verified by

observing a series of appropriate events until we come upon a positive instance; and a single positive case is sufficient to determine that the hypothesis is true. On the other hand, if the class of variables specified in the hypothesis is infinite in size, or at least indefinitely large, we shall never be able to determine that the hypothesis is strictly false. We can only observe a finite number of events. If, in that finite number of events, we do not observe a positive instance of our hypothesis, we may well expect this state of affairs to continue for future observations. But we will never be sure that a negative instance of the hypothesis will not eventually occur. Hence, we can appreciate that this type of hypothesis is also unilaterally verifiable; it is possible to determine that it is strictly true but not possible to determine that it is strictly false. Let us illustrate the inferences involved in testing the existential hypothesis. For the case of a positive evidence report:

Existential Hypothesis:	There is an a such that if a then b	
Positive Evidence Report:	a and b	
	↓	(Deductive Inference)
Conclusion:	Therefore, the hypothesis is (strictly) true.	

For the case of a negative evidence report:

Existential Hypothesis:	There is an a such that if a then b	
Negative Evidence Report:	a and not b	
	↓	(Inductive Inference)
Conclusion:	Therefore, the hypothesis is not confirmed.	

To illustrate by means of our previous example:

Existential Hypothesis:	There is a rat that, if it is reinforced at the end of its maze runs, then it will learn to run that maze with no errors.
Positive Evidence Report:	A group of rats was reinforced at the end of their maze runs and at least one of those rats learned to run that maze with no errors.
Conclusion:	The hypothesis is (strictly) true.

Existential Hypothesis:	There is a rat that, if it is reinforced at the end of its maze runs, then it will learn to run that maze with no errors.
Negative Evidence Report:	A group of rats was reinforced at the end of their maze runs and none of those rats learned to run that maze with no errors.
Conclusion:	The hypothesis is not confirmed.

Limited Hypotheses. Limited hypotheses (offered as direct statements) are always verifiable because it is possible to make a deductive inference from the evidence report (whether positive or negative) to the hypotheses. Thus, a limited hypothesis, such as "if rat number 3 is reinforced at the end of its maze runs, then that rat will learn to run that maze with no errors" is completely (bilaterally) verifiable. For instance, if we find that rat number 3 does learn to run the maze with no errors, we may conclude that the hypothesis is true. But if he does not, we may conclude that the hypothesis is false.

Irrelevant Evidence Reports. One final matter needs to be emphasized: the importance of satisfying the antecedent conditions of the hypothesis in the experimental situation. If this requirement is not satisfied, no inference can be made from the evidence report to the hypothesis. Instead, we may only say that the evidence report is irrelevant to the hypothesis and thus does not constitute a test of the hypothesis. For example, if the experimental group of rats in our hypothetical experiment was not reinforced at the end of its maze runs, then whatever the results with regard to the number of errors made, we cannot say that a test of the hypothesis has occurred.

THE REASON FOR THE "IF... THEN..." FORM

In Chapter 3, we said that universal hypotheses can be viewed as assuming the "If... then..." form. After reading this chapter, the reason for our desire to state them in this form, or at least to know that we could state them in this form if we wanted to, should be apparent. To be more explicit, however, let us note that various inferences can only be made validly if the statements (hypotheses and evidence reports) assume certain forms. For the logical rules that we have borrowed are applicable only to certain forms of statements. By using the conditional form for stating hypotheses, and the conjunctive form for stating evidence reports, we are able to satisfy these rules.

Hence the inferences that we discuss in this chapter are valid, because they conform to the rules of logic. Let us emphasize, however, that strictly speaking, it is not necessary actually to state your hypothesis in the if-then form. All that is required is that, whatever the form in which you state your hypothesis, it is *possible* to restate it in the if-then form. If this is not possible, then it is not possible to make valid inferences from the evidence report to your hypothesis. Similar considerations hold true for the evidence report; in order to perform a valid inference from it to the hypothesis, that evidence report must assume the conjunctive form. If you do not actually state your evidence report in the conjunctive form, however, you need not be concerned *as long as it would be possible to restate it in that form*.

APPENDIX TO CHAPTER 12

Note 1 (from page 308). It is necessary to point out an error that we are perpetuating in this chapter. It concerns the inductive inference that is made from the evidence report to the hypothesis. Let us consider the universal hypothesis. For this case we said that an inductive inference may be made from a positive evidence report to the hypothesis, an inference that results in the conclusion that the hypothesis is confirmed. But this is not quite right. Rather, it is a procedure that is universally used. To understand the error, let us emphasize that the valid inferences of inductive logic are specified by the rules of probability logic (cf. Reichenbach, 1949) and it is not possible to find a rule of this sort in the calculus of probability.

To make the problem more specific, let us say that we seek to test the hypothesis that "If a then b ." Further, assume a positive evidence report was obtained: a and b . Now, we make an inductive inference from that evidence report to the hypothesis and conclude that the hypothesis is confirmed. However, our hypothesis is not the only hypothesis that will also predict the consequent condition b . Rather, there is an unspecifiable number of additional hypotheses which will also have b as a consequent condition — if a then b , if a'' then b , if a''' then b , and so on. When we actually find that b is true, then, we do not know which of the numerous hypotheses that imply b is confirmed. Although it is customary to assume that it is our hypothesis, it may just as well be one of the other possible hypotheses that is confirmed. About the best we can say at this point is that any hypothesis that implies b may be confirmed, providing that its antecedent conditions were present in the experimental situation. And since there are numerous unspecified antecedent conditions present in the experimental situation, we cannot say that only a was present, for a' , a'' , a''' , etc. may also have been present. Unfortunately we cannot here prolong our discussion of this problem. Let us simply note that there are, however, certain inferences in probability logic that would, in principle, satisfy our needs. However, much additional work needs to be accomplished before we can make these more complicated inferences in our everyday research. For more information on the nature of the difficulty and a proposed solution you might refer to McGuigan (1956).

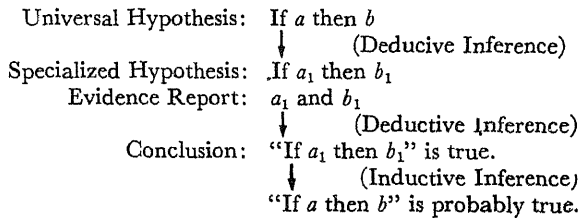
Note 2 (from page 310). At this point, an apparent contradiction in what we have said might occur to you. We previously stated that it is impossible strictly to determine that a hypothesis is true or false. Rather, we must always settle for a probability statement — that the hypothesis has some degree of probability. Here, however, we have said that we can sometimes determine that a hypothesis is false, not merely improbable. How may we reconcile these two statements?

The answer may be stated thus: The deductive inference from the evidence report to the universal hypothesis in effect states that if it is true that the evidence report is negative, then the universal hypothesis is necessarily false. But the evidence report itself rests on a probability statement, for we have formed it by failing to reject the null hypothesis.⁴ Hence, even though we have every reason to believe that the evidence report *really is* negative, we may actually be in error. For this reason we may erroneously conclude that the universal hypothesis is false, even though we are making a deductive inference. Remember that a deductive inference does not lead to an absolutely irrefutable statement about knowledge. Rather, it allows us to say that *if* the premises are true, then it is necessarily the case that the conclusion is true. But the determination of the truth or falsity of the premises (here the evidence report) is still an empirical matter and thus liable to error. It may thus be seen that confirmation and verification are really only different in degree with regard to the final conclusion. Although inductive inferences are used in the former and deductive inferences in the latter, the conclusion regarding the hypothesis still must be evaluated in terms of probability. But the reasons that the conclusion in the two cases is limited to a probability statement are different. In confirmation, the probability character of the conclusion results from the probability of the evidence report *and* the fact that an inductive inference must be made. In verification, however, the probability character of the conclusion rests *only* on the probability character of the evidence report since a deductive (not an inductive) inference is made. In general then, we can say that the conclusion in the case of verification is considerably more probable than that in the case of confirmation. This is so because of the high probability (frequently extremely high) of the evidence report.

Note 3 (from page 310). But this procedure is not quite right. To understand this let us observe that the hypothesis is universal in nature, as we said above, it refers to *all* rats, etc. But we are certainly in no position to test it on all rats; we must be content with a particular sample of rats that we have drawn from a population. We draw a deductive inference that the hypothesis is applicable to our particular sample of rats, i.e., if the hypothesis covers all rats, it certainly covers the particular rats with which we are dealing. By the same token, our evidence report is limited to results obtained on our particular group of rats. But as we have stated, the evidence report (i.e., *a* and *b*) is universal in nature. Thus, "*a* and *b*" states, in effect, that "*all* rats were reinforced at the end of their maze runs and *all* those rats learned to run the maze with no error." But this is not true. We are able to advance

⁴Relationships between the empirical and the null hypotheses may be considered in a far more complicated manner than our simplified (and basic) usage suggests. Refer to Binder (1963), Edwards (1965), Grant (1962), and Wilson, Miller and Lower (1967) for an introduction to a lively controversy on these relationships.

our evidence report for *only* those rats that we studied. Hence, the evidence report is much more specialized or restricted in scope. We should indicate this restriction in some way. This may be accomplished by placing a subscript to the variables contained in the evidence report: a_1 and b_1 , which shall then be read "a (certain) group of rats was reinforced at the end of its maze runs and that group of rats learned to run the maze with no errors." Similarly, the hypothesis that is the result of the deductive inference from the general hypothesis must be stated in specialized form: If a_1 then b_1 . We then confront the specialized hypothesis with the (specialized) evidence report and reach a conclusion (by way of a deductive inference) as to its truth or falsity. And from this conclusion we go back to our universal hypothesis, again by way of an inductive inference, to determine its probability. This general procedure may thus be represented as follows:



In terms of our example, the steps are:

- Universal Hypothesis: If (all) rats are reinforced at the end of their maze runs, then those rats will learn to run that maze with no errors.
- Specialized Hypothesis: If a (specific) group of rats is reinforced at the end of its maze runs, then *those* rats will learn to run that maze with no errors.
- Evidence Report: A (specific) group of rats was reinforced at the end of its maze runs and *those* rats learned to run that maze with no errors.
- Conclusion: a) The specialized hypothesis is true.
b) The universal hypothesis is probably true.

We can now see that there are two reasons that it is an inductive inference that is made from the evidence report to the hypothesis: (1) because the evidence report states what was found for a sample of subjects and the hypothesis is concerned with populations, hence one proceeds from "the specific to the general"; (2) because it is inferred that it is the hypothesis "if a then b " that is confirmed rather than, as discussed in note 1 above, the hypotheses "if a then b ", "if a' then b ", and so forth.

Note 4 (from p. 311). The probability character of empirical hypotheses exerts itself in still other ways, even though they are verifiable, as discussed above. For instance, we may ask how many trials a rat should be run before

we are convinced that he would never learn to run the maze perfectly. Say that after a goodly number of trials the rat has still failed to learn. Yet it may be on just the next trial (no matter where we stopped running him) that he would demonstrate flawless performance. One answer is to be more precise in the statement of our antecedent conditions, i.e., to specify a certain number of trials. For example, "If rats are reinforced at the end of each of 30 maze runs, then. . . ." In this case, if they demonstrate the required performance or if they don't after 30 trials, our conclusion is "certain" and the verifiable nature of the hypothesis is preserved.

THE INDUCTIVE SCHEMA

An Overview of Some Characteristics of Science

"Dr. Watson, Sherlock Holmes," said Stamford introducing us.

"How are you?" he said cordially, gripping my hand with a strength for which I should hardly have given him credit. "You have been in Afghanistan, I perceive."

"How on earth did you know that?" I asked in astonishment . . . "You were told, no doubt."

"Nothing of the sort. I knew you came from Afghanistan. From long habit the train of thoughts ran so swiftly through my mind that I arrived at the conclusion without being conscious of intermediate steps. There were such steps, however. The train of reasoning ran, 'Here is a gentleman of a medical type, but with the air of a military man. Clearly an army doctor, then. He has just come from the tropics, for his face is dark, and that is not the natural tint of his skin, for his wrists are fair. He has undergone hardship and sickness, as his haggard face says clearly. His left arm has been injured. He holds it in a stiff and unnatural manner. Where in the tropics could an English army doctor have seen so much hardship and had his arm wounded? Clearly in Afghanistan.' The whole train of thought did not occupy a second. I then remarked that you came from Afghanistan, and you were astonished." (Doyle, 1938, pp. 6, 14)¹.

¹Reprinted by permission of the Estate of Sir Arthur Conan Doyle.

This, the first meeting between Holmes and Watson, is a relatively simple demonstration of Holmes' ability to reach conclusions that confound and amaze Watson. It serves well to illustrate what Reichenbach has called the *inductive schema*. The reconstruction of Holmes' reasoning is presented in the inductive schema shown in Figure 13.1. The observational information

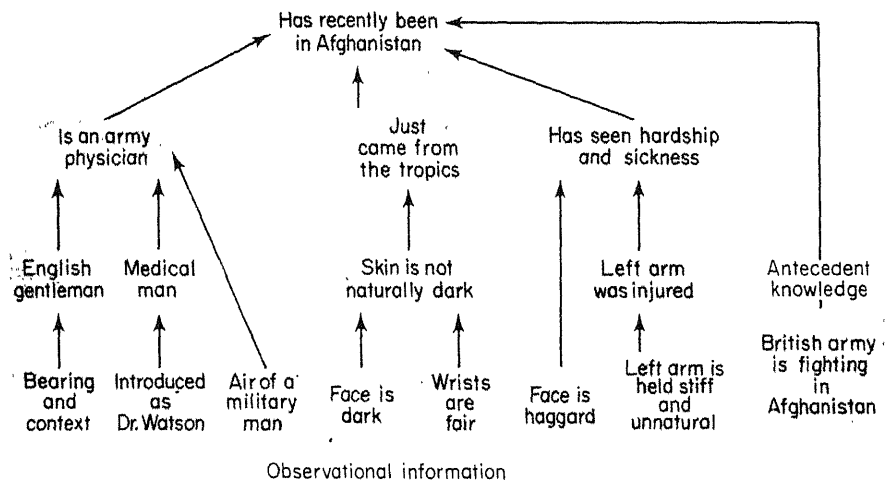


FIGURE 13.1.

An inductive schema based on Sherlock Holmes' first meeting with Dr. Watson.

available to Holmes is presented at the bottom of that schema. On the basis of this information Holmes infers certain intermediate conclusions. For example, he observed that Watson's face was dark, but that his wrists were fair. These two bits of information immediately led to the conclusion that Watson's skin is not naturally dark. He must therefore have recently been in an area where there was considerable sun; Watson had probably "just come from the tropics." From the several intermediate conclusions it was then possible for Holmes to infer the final conclusion, that Watson had just recently been in Afghanistan. You should trace through each step of Holmes' reasoning process, as represented in the inductive schema, to make sure that you understand how it was constructed. You might even want to construct such a schema for yourself from any of Holmes' other amazing processes of reasoning.

Let us now turn to an example from physics, an inductive schema that represents the process of scientific reasoning. In Figure 13.2 we have partially reconstructed the evolution of this science. In the bottom row are some of the basic data (evidence reports of investigation) from which more general statements were made.

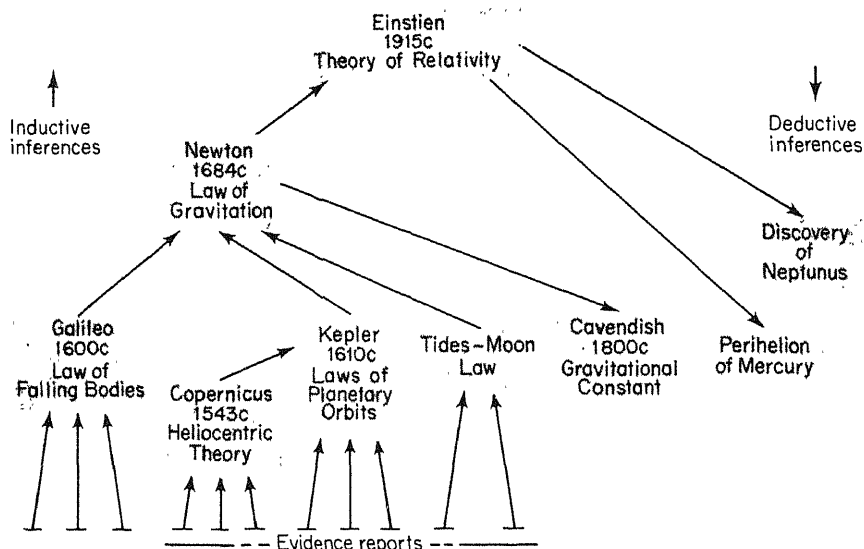


FIGURE 13.2.

An inductive schema which partially represents the development of physics. (After Reichenbach.)

For instance, Galileo conducted some experiments in which he rolled balls down inclined planes. He measured two variables, the time that the bodies were in motion and the distance covered at the end of various periods of time. He found that the distance traveled was related in a specific manner to the amount of time that the bodies were in motion. This relationship is known as the Law of Falling Bodies.²

Copernicus was dissatisfied with the Ptolemaic theory that the sun rotated around the earth, and on the basis of extensive observations and considerable reasoning advanced the Heliocentric (Copernican) Theory of Planetary Motion that the planets rotate around the sun. Kepler based his laws on his own meticulous observations, the observations of others, and on Copernicus' theory. The statement of his three laws of Planetary Orbits (among which was the statement that the earth's orbit is an ellipse) was a considerable advance in our knowledge.

There has always been interest in the height of the tides at various localities,

²More precisely, the Law of Falling Bodies is that $S = \frac{1}{2}gt^2$ where S is the distance the body falls, g the gravitational constant, and t the time that it is in motion. History is somewhat unclear about whether Galileo conducted similar experiments in other situations, but it is said that he also dropped various objects off the Leaning Tower of Pisa and obtained similar measurements.

and it is natural that precise recordings of this phenomenon would have been made at various times during the day. Similar observations were made of the location of the moon. Then, on the basis of these two sets of observations, it was possible to state a relationship, known as the Tides-Moon Law. This relationship states that high tides occur only on the parts of the earth that are nearest to, and farthest from, the moon, respectively. It follows that as the moon moves about the earth, the location of high tides shifts accordingly.

Largely on the basis of the preceding relationships, Newton was able to formulate his law of gravitation. Briefly, this law states that the force of attraction between two bodies varies inversely with the square of the distance between them. As an example of a prediction from a general law, we illustrate (the first downward arrow of Figure 13.2) the prediction of the gravitational constant from Newton's law, the precise determination of which was made by Cavendish.

The crowning achievement of the portion of physics that we are considering came with Einstein's statement of his general theory of relativity. One particularly interesting prediction made from the theory of relativity concerned the perihelion of Mercury. Newton's equations failed to account for a slight discrepancy in Mercury's perihelion. This discrepancy between Newton's equations and the observational findings was accounted for precisely by Einstein's theory. The "chain reaction" of discoveries in science is illustrated by the fact that Einstein's work on the movement of Mercury's perihelion led to the discovery of the planet Neptune by Leverrier.

This brief discussion of the evolution of a portion of physics is, of course, inadequate for a proper understanding of the subject matter involved. Each step in the story constitutes an exciting tale that you might wish to follow up in detail. And where does the story go from here? One of the problems that has been bothering physicists and philosophers is how to reconcile the area of physics depicted in Figure 13.2 with a similar area known as quantum mechanics. To this end physicists such as Einstein and Schrödinger have attempted to develop a "unified field" theory that will encompass Einstein's theory of relativity as well as the principles of quantum mechanics. Consideration of such higher-level principles is beyond our present scope. Our purpose is fulfilled by presenting the general *form* of scientific progress, as shown in Figure 13.2, which is an evolution that may easily be man's crowning intellectual achievement.

We shall now use these two schemata as a basis for considering a number of characteristics of science. Specifically, they will help us (1) to discuss the process of scientific generalization; (2) to elaborate our distinction between inductive and deductive inferences; (3) to understand Reichenbach's concept of concatenation; (4) to understand the nature of explanation; and (5) to understand the nature of prediction.

GENERALIZATION

Galileo conducted a number of specific experiments. Each experiment resulted in a statement that there was a relationship between the distance traveled by balls rolling down an inclined plane and the time that they were in motion. From these specific statements he then advanced to a more general statement: The relationship between distance and time obtained for the bodies in motion was true for *all* falling bodies, at all locations, and at all times.

Copernicus observed the position of the planets relative to the sun. After making a number of specific observations, he was willing to generalize to positions of the planets that he had not observed. The observations that he made fitted the heliocentric theory, that the planets revolved around the sun. He then made the statement that the heliocentric theory held for positions of the planets that he had not observed. And so it is for Kepler's laws and for the Tides-Moon law. In each case a number of specific statements based on observation (evidence reports) were made. Then from these specific statements came a more general statement. It is this process of proceeding from a set of specific statements to a more general statement that is referred to as *generalization*. The general statement, then, includes not only the specific statements that led to it but also a wide variety of other phenomena that have not been observed.

This process of increasing generalization continues as we read up the inductive schema. Thus Newton's law of gravitation is more general than any of those that are lower in the schema. We may say that it generalizes Galileo's, Copernicus', Kepler's, and the Tides-Moon Laws. Newton's law is more general in the sense that it includes these more specific laws and that it makes statements about phenomena that are other than the ones on which it was based. In turn, Einstein formulated principles that were more general than Newton's, principles that included Newton's and therefore all of those lower in the schema.

MORE ON INDUCTIVE AND DEDUCTIVE INFERENCES

In the inductive schema of Figure 13.2, we may observe that inductive inferences are represented when arrows point up, deductive inferences by arrows that point down. In Figure 13.1 we have only inductive inferences, and these are liable to error. For instance, Watson was introduced as "*Dr. Watson*"; on the basis of this information Holmes concluded that Watson was a medical man. Is this necessarily the case? Obviously not, for Watson may have been some other kind of doctor, such as a Doctor of Philosophy. Similarly, consider the observational information, "*left hand held stiff and unnatural*," on the basis of which Holmes concluded that "*the left arm was*

injured." This conclusion does not necessarily follow, since there could be other reasons for the condition (Watson might have been organically deformed at birth). In fact, was it necessarily the case that Watson had just come from Afghanistan? The story may well have gone something like this: Holmes: "You have been in Afghanistan, I perceive." Watson: "Certainly not. I have not been out of London for forty years. Are you out of your mind?"

In a similar vein we may note that Galileo's law was advanced as a general law, asserting that *any* falling body *anywhere at any time* obeyed his law. Is this necessarily true? Obviously not, for perhaps a stone falling off Mount Everest or a hat falling off of a man's head in New York may fall according to a different law than that offered for a set of balls rolling down an inclined plane in Italy many years ago. (We would assume that Galileo's limiting conditions such as that concerning the resistance of air would not be ignored.)

And so it is with the other statements in Figure 13.2. Each conclusion may be in error. As long as inductive inferences are used in such situations, the conclusion will only have a certain degree of probability of being true. Yet, we must continue using inductive inferences. You have no doubt noted that each time a generalization is made, inductive inferences have been used to arrive at that generalization. Since the making of a generalization necessitates saying something about as yet unobserved phenomena, the generalization *must* be susceptible to error.

Let us illustrate deductive logic by referring to our inductive schema for physics. Since Galileo's and Kepler's laws were generalized by Newton's, it follows that they may be deduced from them. In this case, it may be said "If Newton's laws are true, then it is necessarily the case that Galileo's is true, and also that Kepler's are true." Similarly, on the basis of Newton's laws, the gravitational constant was deduced and empirically verified by Cavendish. This deductive inference takes the form: "If Newton's laws are true, then the gravitational constant is such and such." Furthermore, concerning Einstein's principles, we may say: "If Einstein's theory is true, then the previous discrepancy in the perihelion of Mercury may be accounted for."

This discussion allows us to emphasize a very important matter that we have previously mentioned. That is, that a deductive inference does not guarantee us that the conclusion is true. The deductive inference discussed above, for example, does not say that Galileo's law is true. It does say that *if* Newton's laws are true, Galileo's law is true. One may well ask, at this point, how we determine that Newton's laws are true. Or, more generally, how we determine that the premises of a deductive inference are true. The answer, of course, is through the use of inductive logic. For example, we have determined by empirical investigation that the probability of Newton's laws being true is very high. It is, in fact, sufficiently high that we wish to say that it *is* true (in an approximate sense, of course).

CONCATENATION

As we move up the inductive schema we arrive at statements that are increasingly general. And as the generality of a statement increases in the manner depicted in our inductive schema, there is a certain increase in the probability of the statement being true. This increase in probability is the result of two factors. *First*, since the more general statement rests on a wider variety of evidence, it usually has been confirmed to a greater degree than has a less general statement. For example, there is a certain addition to the probability of Newton's law of gravitation that is not present for Galileo's law of falling bodies, since the former is based on inductions from a wider scope of phenomena. *Second*, the more general statement is *concatenated* with other general statements. By concatenated we mean that the statement is "chained together" with other statements and is thus consistent with these other statements. For example, Galileo's law of falling bodies is not concatenated with other statements, and Newton's is. The fact that Newton's law is linked with other statements gives it an increment of probability that cannot be said of Galileo's. We may say that the probability of the whole system being true is greater than the sum of the probabilities of each statement taken separately. It is the compatibility of the whole system, and the support gained from the concatenation that provides the added likelihood.

It also follows that when each individual generalization in the system is confirmed, the entire system gains increased credence. For instance, if Einstein's theory was based entirely on his own observations, and those which it stimulated, its probability would be much lower than it actually is, considering that it is also based on all of the lower generalizations in Figure 13.2. Or let us put the matter another way. Suppose that a new and extensive test determined that Galileo's laws were false. This would mean the complete "downfall" of Galileo's laws, but it would only slightly reduce the probability of Einstein's theory since there is a wide variety of additional confirming data for the latter.

EXPLANATION

The concept of explanation as used in science is sometimes difficult for students to understand, probably because of the common sense use of the term to which they have previously been exposed. One of the common sense "meanings" of the term concerns familiarity. Suppose that you learn about a scientific phenomenon that is new to you. You want it explained; you want to know "why" it is so. This desire on your part is a psychological phenomenon, a motive. When somebody can relate the scientific phenomenon to something that is already familiar to you, your psychological motive is satisfied. You feel as if you understand the phenomenon because of its asso-

ciation with knowledge that is familiar to you. A metaphor is frequently used for this purpose. At a very elementary level, for example, it might be said that the splitting of an atom is like shooting an incendiary bullet into a bag of gunpowder.

Any satisfaction of your motive to relate a new phenomenon to a familiar phenomenon is far from an explanation of it. Explanation is the placing of a statement within the context of a more general statement. If we are able to show that a specific statement belongs in the category of a more general statement, we may say that the specific statement has been explained. To establish this relationship we must show that the specific statement may be logically deduced from the more general statement. To return to a previous example, we might ask how to explain the statement that "John Jones/is anxious." The answer is that this statement can be logically deduced from the more general statement that "All men who bite their fingernails are anxious," and that "John Jones bites his fingernails." This immediately brings us face to face with a matter that we have approached previously from different angles: that such an explanation is accomplished on the assumption that the more general statement is true. Hence, to be more complete, we need to say: "If it is true that 'all men who bite their fingernails are anxious,' and if it is true that 'John Jones is a man who bites his fingernails,' then it is true that 'John Jones is anxious.'" By so deductively inferring this conclusion, we have explained why John Jones is anxious; we have logically deduced that specific statement from the more general statement. Of course, we immediately want to go on: Why are all men anxious? But this is outside our scope, except that we may note that such an explanation would be accomplished by deducing that statement from a still more general one.

Referring to Figure 13.2 we can see that Kepler's laws are more general than the Copernican theory. And since the latter is included in the former, it may be logically deduced from it — Kepler's laws explain the Copernican theory. In turn, Newton's law, being more general than Galileo's, Kepler's, and the Tides-Moon laws, explains these more specific laws; they may all be logically deduced from Newton's law. And finally, all of the lower generalizations may be deduced from Einstein's theory, and we may therefore say that Einstein's theory explains all of the lower generalizations.

PREDICTION

To make a prediction we apply a generalization to a situation that has not yet been studied. The generalization says that all of something has a certain characteristic. When we extend the generalization to the new situation, we are simply saying that the new situation should have the characteristic specified in the generalization. In its simplest form this is what a prediction is, and we have illustrated three predictions in Figure 13.2, the

gravitational constant, the perihelion of Mercury, and discovery of Neptunus. Whether or not the prediction is confirmed, of course, is quite important for the generalization. For if the prediction is confirmed, the probability of the generalization is considerably increased. If it is not confirmed, however, (assuming the evidence report is true and that the deduction is valid), then either the probability of the generalization is decreased, or the generalization must be restricted so that it does not apply to the type of phenomena with which the prediction was concerned.

As an illustration of a prediction let us say that a certain hypothesis was formulated about the behavior of school children in the fourth grade. It was then tested on these children and found to be probably true. The experimenter may wonder whether this hypothesis is also true for all school children. If he thinks so he might generalize his hypothesis to make it applicable to all school children. From such a generalized hypothesis it is possible to derive specific statements concerning any given school grade. For example, he could deductively derive the conclusion that the hypothesis is applicable to the behavior of school children in the fifth grade, in which case he is making a prediction about children of that grade level. He is, thus, making a prediction that his hypothesis is applicable to a novel situation.

We have, in this chapter, attempted to present some of the important characteristics and processes of science in general. With an understanding of these principles we can now turn to a consideration of their role in psychology.

GENERALIZATION, EXPLANATION, AND PREDICTION IN EXPERIMENTATION

In the last chapter we discussed generalization, explanation, and prediction, as well as several other topics, in a rather general way. It now remains for us to consider these topics as they fit into the day to day work of the experimenter. We shall want to discuss some of the mechanics that one might use in these, the final phases of experimentation. The three questions to which we now turn are: How and what does the experimenter generalize? How does he explain his results? And how does he predict to other situations?

GENERALIZATION

A distinction is frequently made between what is known as applied science (technology) and basic or pure science. In applied science the investigator attempts to solve some relatively limited problem, whereas in basic science he attempts to arrive at a general principle. The answer that the applied scientist obtains will usually be applicable only under the specific conditions

of his experiment. The basic scientist's results, however, are likely to be more widely applicable.

An applied psychologist might be called in by the Burpo Company to find out why their soft-drink sales in Atlanta, Georgia were below normal for the month of December. The basic scientist, on the other hand, would be more likely to study the problem of the general relationship between temperature and consumption of liquids. The applied psychologist, in this example, might find that Atlanta was unseasonably cold during December and hypothesize that it was for that reason that sales declined. The basic scientist however, might conclude his research with the more general finding that the amount of liquid consumed by humans depends on the temperature — the lower the temperature, the less they consume. Thus, the latter finding would account for the specific phenomenon in Atlanta, as well as a wide variety of additional phenomena; it is by far the more general in scope.

Although both kinds of research are important, we shall limit our considerations to basic research. Our immediate goal shall be to see how the experimenter arrives at general statements, rather than only specific statements about the results of his research. We shall start our discussion with the assumption that the experimenter wants to generalize his results as widely as is reasonable. Hence we shall consider two questions in detail: By what procedures does he generalize his results; and how wide is "reasonable"?

THE MECHANICS OF GENERALIZATION

Let us say that an experimenter has selected twenty subjects for an experiment. It should be apparent that he is not interested in these twenty subjects in and for themselves but only insofar as they are typical of a larger group. Whatever he finds out about these subjects he assumes will be true for the larger group. In short, he wishes to *generalize* from his twenty subjects to the larger group of subjects. The terms we have previously used in this connection are "sample" and "population." As we said, an experimenter defines a population of subjects that he wishes to make statements about. This population is usually quite large, such as all the students in the university, all dogs of a certain species, or perhaps even all humans. Since it is not feasible to study all the members of such large populations, the experimenter randomly selects a sample therefrom. And since that sample has been randomly selected from the population, it should be representative of that population. Therefore, the experimenter is able to say that what is probably true for the sample is also probably true for the population; he generalizes from the sample of subjects to the entire population of subjects from which they came.¹

¹Even though this statement offers the general idea, it is not quite accurate. If we were to follow this procedure, we would determine that the mean of a sample is, say, 10.32, and generalize to the population, inferring that its mean is also 10.32. Strictly speaking,

The *most important* requirement for generalizing from a sample, is that the sample must be *representative* of the population. The technique that we are using for obtaining representativeness is randomization; if the sample has been randomly drawn from the population, it is reasonable to assume that it is representative of the population. *Only when the sample is representative of the population are we able to generalize from the sample to the population.* We are emphasizing this point to a great extent for two reasons: because of its great importance in generalizing to populations of *subjects*, and because we ourselves want to state a generalization. We want to generalize from what we have said about subject populations to a wide variety of other populations.² For when you conduct an experiment you actually have a number of populations, in addition to the population of subjects, to which you might generalize.

To illustrate, suppose you are conducting an experiment on knowledge of results. You take two groups of subjects and assign them to two conditions: One group receives knowledge of results, and the second (control) group doesn't. The classic task used in studying this problem is line drawing. Subjects are blindfolded and asked to draw five-inch lines. The knowledge-of-results group would be told whether their lines were too long, too short, or correct, while the control group would be given no knowledge about the lengths of their lines. We are dealing with several populations: of subjects, of experimenters, of tasks, and of various stimulus conditions. Since we wish to generalize to a population of subjects, we randomly draw our sample from that population and randomly assign them to the two groups. If we find, as we certainly should, that the knowledge-of-results group performs better than the control group, we can safely say that this is probably also true for the population of subjects.

But what about the experimenter? We have controlled this variable, presumably, by having a single experimenter handle all the subjects. But can we say that the knowledge-of-results group will always be superior to the control group *regardless of who is the experimenter?* In short, can we generalize from the results obtained by our single experimenter to all experimenters? This question is difficult to answer. Let us imagine a population of experimenters, made up of all people who conduct experiments. Strictly speaking, then, we should take a random sample from the population of experimenters and have each member of our sample conduct the experiment for himself. Suppose that we define our population of experimenters in such a way that it

this procedure is not reasonable, for it could be shown that the probability of such an inference is .00. A more suitable procedure is known as "confidence interval estimation," whereby one infers that the mean of the population is "close to" that for the sample. Hence, the more appropriate inference might be that, on the basis of a sample mean of 10.32, the population mean is between 10.10 and 10.54.

²For an elaboration of matters relating to generalization to nonsubject populations you might refer to Brunswick (1956) and Hammond (1948, 1954).

includes 500 people and that we randomly select a sample of ten experimenters from that population. Further assume that we have selected a sample of 100 subjects. We would then randomly assign the 100 subjects to two groups, then we would randomly assign five subjects in each group to each experimenter. In effect, then, we will repeat our experiment ten times. We have now not only controlled the experimenter variable by balancing, but we have also sampled from a population of experimenters. Assume that the results come out approximately the same for each experimenter — that the performance of the knowledge-of-results subjects is about equally superior to their corresponding controls for all ten experimenters. In this case we are able to generalize the results to the population of experimenters as follows: For the population of experimenters sampled (and also for the population of subjects sampled), providing knowledge of results under the conditions of this experiment leads to performance that is superior to that derived from not providing knowledge of results.

By “under the conditions of this experiment” we mean two things: with the specific task used, and under the specific stimulus conditions that were present for the subjects. Concerning the first, our question is this: Since we found that the knowledge-of-results group was superior to the control group on a line-drawing task, would that group also be superior in learning other tasks? Of course, the answer is that we do not know from this experiment. Consider a population of *all* the tasks that human subjects could learn, such as doing line drawing, learning Morse code, hitting a golf ball, assembling parts of a radio, and so forth. If we wish to make a statement about the effectiveness of knowledge of results for all tasks, then, as before, we must obtain a representative sample from that population of tasks. By selecting only a line-drawing task we did not do this and therefore cannot generalize back to the larger population of tasks. The proper procedure to generalize to all tasks would be randomly to select a number of tasks from that population. We would then conduct the same experiment for each of those tasks. If we find that on each task studied the knowledge-of-results group is superior to the control group, then we can say that for all tasks, knowledge of results leads to performance that is superior to that gained from a lack of knowledge of results.

Now what about the various stimulus conditions that were present for our subjects? For one, they were blindfolded. But there are different techniques for “blindfolding” subjects. One experimenter might use a large handkerchief, another might use opaque glasses, and still another might place a large screen between the subject’s eyes and his hands so that although the subject would be able to see, he could not view the length of his lines. Would the knowledge-of-results condition be superior to the control condition regardless of the technique of blindfolding? What about other stimulus conditions?

Would the specific temperature be relevant? How about the noise level? And so on — one can conceive of a number of populations of these stimulus conditions. Strictly speaking, if an experimenter wishes to generalize to the populations of stimuli present, he should randomly sample from those populations. Take temperature as an example. If he wishes to generalize his results to all reasonable values of this variable, then he should randomly select a number of temperatures. He would then repeat his experiment for each temperature value studied. If he finds that regardless of the temperature value studied the knowledge-of-results condition is always superior, he can generalize his findings to the population of temperatures sampled. Only by systematically sampling the various stimulus populations can the experimenter, strictly speaking, generalize his results to those populations.

At this point it might appear that the successful conduct of psychological experimentation is hopelessly complicated. One of the most discouraging features of psychological research is the difficulty encountered in confirming the results of previous experiments. When one experimenter (Jones) finds that variable A affects variable B, all too frequently another experimenter (Smith) achieves different results. The reason for this lack of repeatability was discussed in Chapter 6, on control, and Chapter 10, on factorial designs. Looking at it from the present point of view, we might explain the differences in findings by the fact that Experimenter Jones held a number of conditions constant in his experiment and then generalized to the populations of these conditions. For example, he may have held the experimenter variable constant, and at least implicitly generalized to a population of experimenters. Strictly speaking, he should not have done that, for he did not randomly sample from a population of experimenters. Let us then assume that his generalization was in error and that the results he obtained are valid only when *he* is the experimenter. If this is the case, then a different set of results may very well be obtained with a different experimenter.

Psychological research (or *any* research for that matter) frequently does become discouraging. After all, if it were easy there would be little joy to it. The toughest nut to crack yields the tastiest meat. Psychologists, however, are beginning to systematically investigate experimental situations more thoroughly than in the past. They are increasingly seeking to account for conflicting results in independent experiments. This is one of the reasons that factorial designs are being more widely used, for they are wonderful devices for sampling a number of populations simultaneously. To illustrate, suppose that we wish to generalize our results to populations of subjects, experimenters, tasks, and temperature conditions. We could conduct several experiments here, but let us say that we conduct only one experiment using four independent variables, each varied in the following ways: (1) knowledge of results, two ways (knowledge and no knowledge); (2) experimenters

varied in six ways; (3) tasks varied in five ways; and (4) temperature varied in four ways. Assume that we have chosen the values of the last three variables at random. The resulting factorial is presented in Table 14.1.

It can thus be seen that we have a $6 \times 5 \times 4 \times 2$ factorial design. What if we find a significant difference for the knowledge of results variable, but no significant interactions? In this case we could rather safely generalize about knowledge of results to our experimenter population, to our task population, to our temperature population, and also, of course, to our subject population.

Table 14.1. *A $6 \times 5 \times 4 \times 2$ factorial design for studying the effect of knowledge of results when randomly sampling from populations of experimenters, tasks, temperatures, and subjects.*

		Knowledge of results						No knowledge of results					
		Experimenters						Experimenters					
		#1	#2	#3	#4	#5	#6	#1	#2	#3	#4	#5	#6
Temperature 4	Tasks												
	D												
	C												
	B												
Temperature 3	Tasks												
	D												
	C												
	B												
Temperature 2	Tasks												
	D												
	C												
	B												
Temperature 1	Tasks												
	D												
	C												
	B												

At this point it is well to recall our discussion from Chapter 10 on factorial designs. There we distinguished between the case of a fixed model and the case of a random model. For the case of a fixed model, we said that the

experimenter selects the values of his independent variables for some specific reason; he does not randomly select them from a population. For the case of a random model, however, he defines his population and then randomly selects values from that population. The relevance of that distinction should be apparent, for only in the case of random variables can you safely generalize to the population. If you select the values of your variables in a nonrandom fashion, any conclusions must be restricted to those values. Let us illustrate by considering the temperature variable again. Suppose that we are particularly interested in three specific values of this variable, 60 degrees, 70 degrees, and 80 degrees. Now, whatever our results, they will be limited to those particular temperature values. On the other hand, if we are interested in generalizing to all temperatures between 40 and 105, we would write each number between 40 and 105 on a piece of paper, place all these numbers in a hat, and draw several values from the hat. Then whatever the experimental results we obtain, we can safely generalize back to that population of values, for we have randomly selected our values from it.³

THE LIMITATION OF GENERALIZATIONS

How widely is it reasonable to generalize? Let's say that we are interested in whether Method A or Method B of learning leads to superior performance. Assume that one experimenter tested these methods on a sample of college students and found Method A to be superior. He unhesitatingly generalizes his results to all college students. Another experimenter becomes interested in the problem and repeats the experiment. He finds that Method B is superior. We wish to resolve the contradiction. After studying the two experiments we may find that the first experimenter was in a women's college, whereas the second was in a men's college. A possible reason for the different results is now apparent. The first experimenter generalized to a population of male and female students without randomly sampling from the former (as also did the second, but without sampling females). To determine whether we have correctly ascertained the reason for the conflicting results we design a 2×2 factorial experiment in which our first variable is methods of learning, varied in two ways, and our second is sex, varied, of course, in two ways. We randomly draw a sample of males and females from a college population. Assume that our results come out with the following mean values, where the higher the score, the better the performance (Table 14.2).

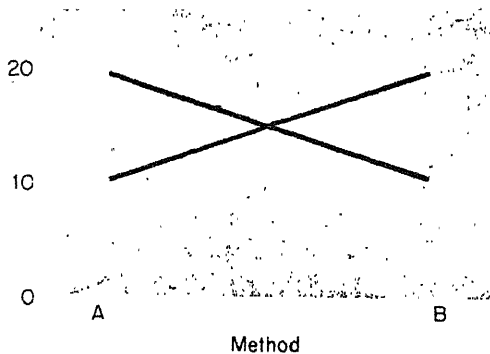
Graphing these results, we can clearly see that an interaction exists between sex and methods such that females are superior with Method A and males

³Assuming, of course, that we select enough values to study. Just as with sampling from a subject population, the larger the number of values selected, the more likely that the sample is representative of the population.

Table 14.2. *A 2 × 2 Factorial Design with Fictitious Means.*

		Methods	
		A	B
Sex	Males	10	20
	Females	20	12

are superior with Method B (Figure 14.1). We have thus confirmed the results of the first experimenter in that we found, as did he, that Method A is superior for females; similarly, we have confirmed the results of the second experimenter since we found that Method B is superior for males. We have,

**FIGURE 14.1.**

Indicating an interaction between methods of learning and sex.

therefore, established the reason for conflicting results. But we cannot make a simple statement about the superiority of a method that generally applies to everybody; the discovery of this interaction limits the extent to which a simple generalization can be offered.

This example provides us with the opportunity to consider the matter of limitations of our generalizations in a broader fashion. Our goal, we have said, is to attempt to make statements of as great a degree of generality as we can. And we would like those statements to be as simple (parsimonious) as

possible (see p. 49). Unfortunately, though, nature does not always oblige us and, to make general statements, we often must complicate them in order to accurately describe events that we study. We can, in short, expect to find a large number of interactions with our experimental treatments, and we should explicitly design our experiments so that we can establish interactions where in fact they exist. The alternative of failing to look for interactions amounts to blinding ourselves to truth, with the consequence that we have difficulty in confirming previous experimentation.

In general, then, the experimenter should systematically study variables that might interact with his variables of primary interest. It is often very easy to construct an experimental design so that such interactions can be studied. The example given above is a case in point, for one can conveniently analyze his results as a function of gender, or other subject characteristic such as anxiety. We have emphasized the importance of controlling the experimenter variable and have shown that this variable may exert subtle influences that may affect dependent variable measures (see also, Engram, 1966). Where more than one experimenter collects data in a given experiment (and this happens in about 48 per cent of published experiments, as reported by Woods, 1961) it is "a natural" to analyze the results as a function of experimenters in order to see if this variable interacts with that of primary interest. Similar variables that may be built into a factorial design for this purpose might be environmental temperature, type of task, nature of equipment used (e.g., memory drum X as against memory drum Y), and so forth.

Let us now examine more closely the possible outcomes with regard to the variable of secondary interest. We shall consider three possible cases. Assume that we vary the independent variable of primary interest in two ways so that we have what we conventionally call an experimental and a control group. The variable of secondary interest may be varied in several ways but, for the moment, let us vary it in only two ways. For instance, let us say that two experimenters collect data in a two-randomized groups design so that we can analyze the data as a 2×2 factorial design (see Table 14.3). The three possible outcomes are discussed as the following cases.

Table 14.3. *A Two-Groups Design in which the Data are Analyzed as a Function of Two Experimenters.*

		Independent variable	
		Experimental	Control
Experimenter	#1		
	#2		

Case I. This case occurs when Experimenters 1 and 2 obtain precisely the same results. The results are graphed in Figure 14.2 where we can note that the lines are parallel. In this instance the variable of secondary interest does not influence the dependent variable measure and we would conclude that it

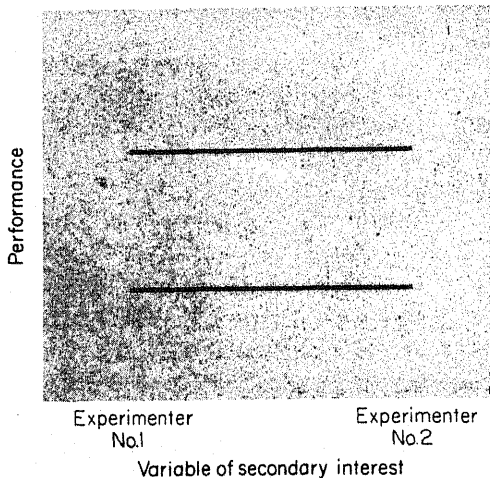


FIGURE 14.2.

Population values for Case I showing no interaction between the variables of secondary interest (e.g., experimenters) and treatments. Furthermore there is no differential effect for varying the variable of secondary interest on dependent variable scores.

does not interact with our variable of primary interest. In this case a difference between our experimental and control groups can be generalized with regard to the variable of secondary interest. There is but one remaining point: We could not possibly have known this unless we had designed and analyzed our experiment to find it out.

As an empirical illustration of Case I, consider an experiment involving two methods of learning and three data collectors (McGuigan, Hutchens, Eason and Reynolds, 1961). An analysis of variance indicated that there was a significant difference between methods but that the differences among experimenters, and the experimenter \times methods interactions were not significant. A graph of the dependent variable scores for the two methods as a function of experimenters is presented in Fig. 14.3. Lines of best fit for these sample points do not deviate significantly from horizontal lines. Hence, we have some reason for generalizing the methods results to a population of experimenters, though of course a larger sample of this population would be preferred.

Case II. The second general possibility is that variation of the variable of secondary interest *does* affect the dependent variable, but it affects all subjects

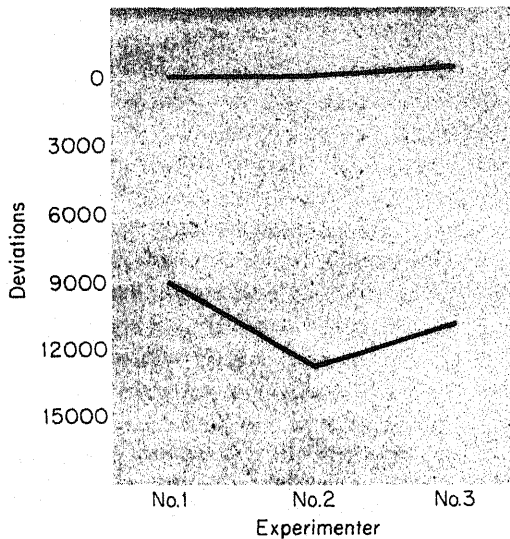


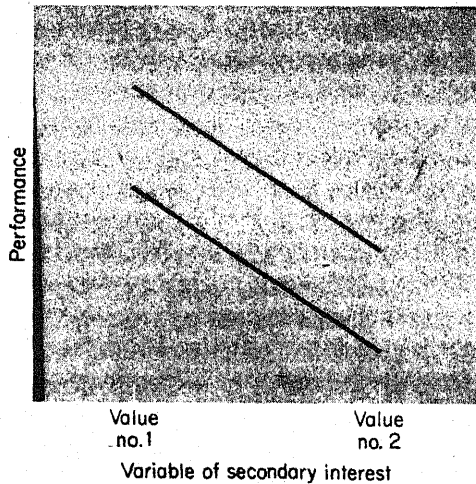
FIGURE 14.3.

Sample values illustrating Case I. Three experimenters and two methods (knowledge of results) were used. The interaction between experimenters and methods is not significant.

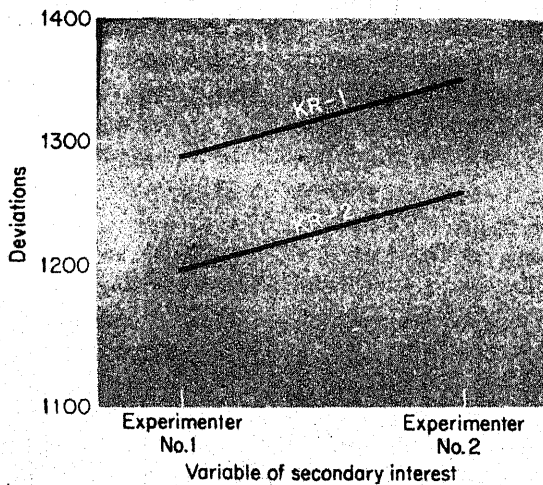
in the same way, regardless of the experimental condition to which those subjects were assigned. For example, we might suppose that subjects assigned to Experimenter 1 (or temperature A, or task X) perform at a higher level on the average than do those subjects assigned to Experimenter 2 (or temperature B, or task Y). But the experimental group is equally superior to the control group for both experimenters, or what have you. This case is illustrated in Figure 14.4.

In Figure 14.5 we have selected values from an experiment in which there was a significant difference between experimenters, but lack of an interaction between experimenters and methods as an empirical illustration of Case II (McGuigan, 1959). Since in Case II we are able to reach the same conclusion with regard to our hypothesis regardless of which experimenter conducted the experiment, we are not immediately interested in the experimenter difference and have a basis for generalizing the results with regard to methods to that population. There is, in short, a lack of interaction that could limit our generalization. As an adjunct to this case, however, we note that a particular kind of behavior is influenced by this secondary variable, information that may be valuable for further experimentation.

Case III. In Cases I and II we have justification for generalizing to the population of the secondary variable to the extent to which that population has been sampled. In Case III, however, we must deal with an interaction.

**FIGURE 14.4.**

Population values for Case II showing no interaction between methods and the variable of secondary interest. But the variable of secondary interest does differentially affect the dependent variable scores.

**FIGURE 14.5.**

Sample values illustrating Case II. The results are for two methods of presenting knowledge of results.

To take an extreme example, suppose that the control group is superior to the experimental group for one experimenter but that the reverse is the case for the second experimenter (Figure 14.6). In this event the extent to which we can generalize to a population is sharply restricted, particularly since we probably don't know the precise ways in which the two experimenters differ.

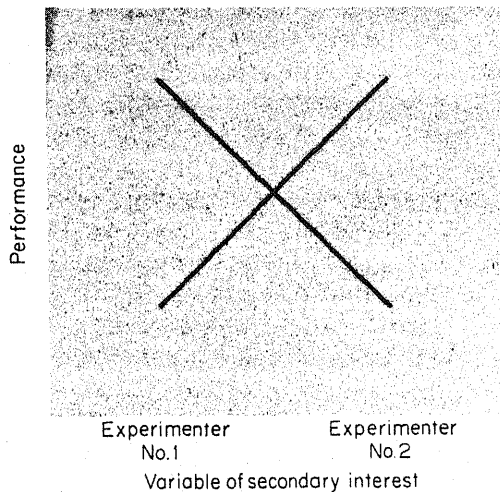


FIGURE 14.6.

Population values for Case III showing one possible interaction between a variable of secondary interest (here experimenters) and treatments.

To understate the matter, the discovery of an interaction of this sort tells us to proceed with caution. It is, however, heartening to note that psychologists are now rather vigorously investigating the effects of the experimenter (and other secondary variables) on their data. Let us briefly look at two interesting studies in which interactions with experimenters have been established.

The first was a verbal conditioning study using the response class of hostile words emitted in sentences (Binder, McConnell, and Sjöholm, 1957). Whenever the subject used a hostile word in a sentence the experimenter reinforced that response by saying "good." Two groups were used, a different experimenter for each group. The two experimenters differed in gender, height, weight, age, appearance, and personality:

"The first . . . was . . . an attractive, soft-spoken, reserved young lady . . . 5'½" in height, and 90 pounds in weight. The . . . second . . . was very masculine, 6'5" tall, 220 pounds in weight, and had many of the unrestrained personality characteristics which might be expected of a former marine captain — perhaps more important than their actual age difference of about 12 years was the difference in their age appearance: The young lady could have passed for a high school sophomore while the male experimenter was often mistaken for a faculty member" (Binder et al., 1957, p. 309).

The results of this experiment are presented in Figure 14.7. First we may note that since the number of hostile words emitted by both groups increases as number of trials increase, the subjects of both experimenters were successfully conditioned. During the first two blocks of learning trials, however, the

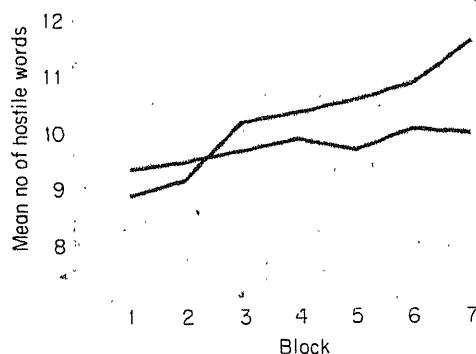


FIGURE 14.7.

Learning curves for two groups treated the same, but with different experimenters. The steeper slope for the subjects of the female experimenter illustrates an interaction between experimenters and stage of learning (after Binder *et al.*, 1957).

subjects of the female experimenter were inferior to those of the male experimenter. On succeeding blocks the reverse is the case, and the two curves intersect. In short, there is an interaction between experimenters and learning trials such that the slope of the learning curve for the female experimenter is steeper than that for the male experimenter. If we, therefore, wish to offer a generalization about the characteristics of the learning curve, it must be tempered by considering the nature of the experimenter. Exactly why this difference occurred is not clear, but we may speculate with the authors that the female experimenter "provided a less threatening environment, and the *S*'s consequently were less inhibited in the tendency to increase their frequency of usage of hostile words" (Binder, *et al.*, 1957, p. 313). Presumably some reverse effect was present early in learning.

In the second example of Case III Spire (1960) selected a group of subjects who scored high on the Hysteria Scale of the Minnesota Multiphasic Personality Inventory and a second who scored high on the Psychasthenic Scale. The subjects were then given one of two sets when they entered the experimental situation: for the positive set the subject was told that the experimenter was a "warm, friendly person, and you should get along very well"; for the negative set the subject was told that the experimenter may "irritate him a bit, that he's not very friendly, in fact kind of cold." Even though the groups of subjects had different sets for the experimenter, the same person was, in fact, the experimenter for both groups.

The subjects were then conditioned to emit a class of pronouns that was reinforced by saying "good." An analysis of variance of Spire's results

indicated that there was a significant difference between positive and negative sets for the experimenter such that subjects with the positive set conditioned better than those with a negative set. Furthermore, and this is the point of present interest, there was a significant interaction between set for the experimenter and personality of the subject (whether he was an hysteric or a psychasthenic). To illustrate this interaction we have plotted the terminal conditioning scores under these four conditions in Figure 14.8. We can thus

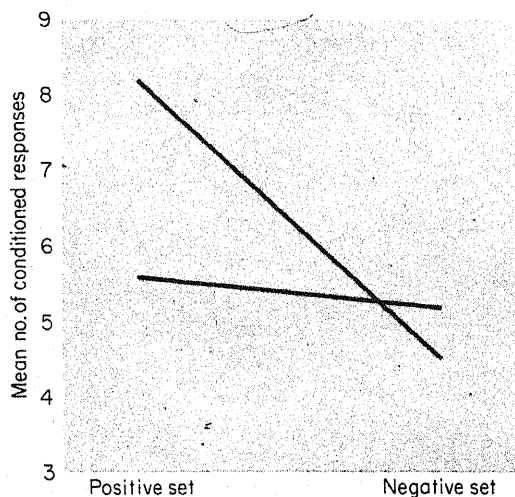


FIGURE 14.8.

An interaction between set for the experimenter and personality characteristic of subjects. The effect of set depends on whether subjects are hysterics or psychasthenics (after Spires, 1960).

see that the hysterics who were given a positive set had higher scores than those given a negative set. There is, though, little difference between the two groups of psychasthenics.

This type of research is especially valuable to us because of its analytic nature; it suggests, for instance, that we can generalize conditioning results with regard to this particular experimenter characteristic for one kind of subject, but not for another kind. Continuous and sustained analysis of the various secondary variables in an experimental situation will eventually allow us to advance our generalizations so that we can have great confidence in their confirmability.

Having now specified these three cases, let us attempt to summarize where we stand.

First, if you have not sampled from a population of some secondary variable, you should, strictly speaking, not generalize to that population. If, for

instance, there is but one data collector for your experiment the best that you can do is to attempt to hold his influence on the subjects constant. If, however, you have previous knowledge that no interaction has been found between your independent variable and the populations for other variables to which you wish to generalize, then your generalization to those populations will probably be valid.

Second, if you have systematically varied some variable of secondary interest to you, then you should investigate the possibility that an interaction exists between it and your variable of primary interest. If, for instance, more than one data collector has been used you should: (1) specify techniques for controlling this variable (see p. 139); (2) analyze and report your data as a function of experimenters; and (3) test for interactions between experimenters and treatments. Should your analysis indicate that the experiment is an instance of Cases I or II, the results are generalizable to a population to the extent to which that population has been sampled.⁴ We grant that completely satisfactory sampling of secondary variables can seldom occur, but at least some sampling is better than none. And it is beneficial to *know* and be able to *state* that, within those limitations, the results appear to be instances of Cases I or II.

Third, if you find that your data are an instance of Case III, then you cannot offer a simple generalization for them. If your variable of secondary interest is an operationally defined one, then your generalization can be quite precise, if a bit complicated. Consider as an example the experiment by Reynolds and Meeker (1966) in which the effects of shock and thiosemicarbazide were studied on the development of gastric ulcers in rats. Their generalization is quite exact, and incidentally theoretically very interesting, precisely because the variables interact. They found that subjects who received thiosemicarbazide and shock developed fewer ulcers than those who received shock only, thiosemicarbazide only, or nothing. Hence, there is an interaction such that the combination of the drug and shock led to an increased resistance to stress (ulcers). This is a generalization incorporating two variables, and it is therefore not as simple as those involving only one variable. If, on the other hand, you cannot adequately specify the ways in which your secondary variable differ (as in the case of different data collectors), the extent to which you can generalize is sharply limited. You can only say, for instance, that Method A will be superior to Method B when experimenters similar to Experimenter 1 are used, but that the reverse is the case when experimenters similar to Experimenter 2 are used. This knowledge is of course valuable, but only in a negative sense since we do not know what the different characteristics of the two experimenters are; an interaction of

⁴We are assuming that a random model is used (see p. 272).

this kind tells us to proceed with considerable caution (cf., McGuigan, 1963).

This all may sound rather demanding, and rather than conclude this topic on such a note let us return to the most typical situation that you are likely to face in your elementary work, namely that specified in the first point above.

If you have no knowledge about interactions between your independent variable and the populations to which you wish to generalize, then it is possible to tentatively offer your generalization. Other experimenters may then repeat your experiment in their own laboratories. This implies that the various extraneous variables will assume different values from those that occurred in your experiment (either as the result of intentional control or because they were allowed to randomly vary). If, in the repetitions of your experiment your results are confirmed, it is likely that the populations to which you have generalized do not interact with your independent variable. On the other hand, if repetitions of your experiment by others, with differences in tasks, stimulus conditions, and other factors do not confirm your findings, then there is probably at least one interaction that needs to be discovered. At this point thorough and piecemeal analysis of the differences between your experiment and the repetitions of it needs to take place in order to discover the interactions. Such an analysis might assume the form of a factorial design such as that diagrammed in Table 14.2 and illustrated by Figure 14.1.

This last point leads us to consider an interesting proposal. Some experimenters have suggested that we should use a highly standardized experimental situation for studying particular types of behavior. All experimenters who are studying a given type of behavior should use the same values of a number of extraneous variables — lighting, temperature, noise, and so forth. In this way we can exercise better control and be more likely to confirm each other's findings. The Skinner Box is a good example of an attempt to introduce standardized values of extraneous variables into an experiment, for in the Skinner Box the lighting is controlled (the box is opaque), the noise level is controlled (it is sound deadened), and a variety of other external stimuli are prevented from entering the box. On the other hand, under such highly standardized conditions the extent to which we can generalize our findings is sharply limited. If we continue to proceed in the direction we are now going with each experimenter having different values of his extraneous variables, then when experimental findings are confirmed, we can be rather sure that interactions do not exist. And when findings are not confirmed, we know that we have interactions present that limit our generalizations, and hence we have to initiate experimentation in order to discover them. Regardless of your opinion on these two positions, that in favor of standardization or that opposed, the matter is probably only academic. Whether because scientists cherish their freedom to establish whatever experimental

conditions they want, or because of their lack of concern, it is unlikely that much in the way of standardization will be accomplished in the foreseeable future.

The final matters for us to discuss in this chapter are those of explanation and prediction, particularly as these are relevant to the day-to-day work of the experimenter. First we shall consider explanation, for after that it will be possible to cover the topic of prediction very briefly.

EXPLANATION⁵

The question "why" lies at the heart of scientific investigation.⁶ We have seen that in science "why" amounts to asking: According to what general statements (laws) does a particular event occur? Thus by placing the particular event within the context of a more general statement we can say that the particular event is explained. Let us now consider this process in greater detail.

Hempel and Oppenheim (1948) have offered an example wherein a mercury thermometer is rapidly immersed in hot water. A temporary drop of the mercury column occurs, after which the column rises swiftly. Why does this occur? That is, how might we explain it? Since the increase in temperature affects at first only the glass tube of the thermometer, the tube expands and thus provides a larger space for the mercury inside. To fill this larger space, of course, the mercury level drops. But as soon as the increase in heat is conducted through the glass tube and reaches the mercury, the mercury also expands. And since mercury expands more than does glass (i.e., the coefficient of expansion of mercury is greater than that of glass), the level of the mercury rises.

Now this account, as Hempel and Oppenheim point out, consists of two kinds of statements. Some statements indicate certain conditions that exist before the phenomenon to be explained occurs. These conditions may be referred to as *antecedent conditions* and they include the fact that the thermometer consists of a glass tube that is partly filled with mercury, that it is immersed in hot water, and so on.⁷ The second kind of statement expresses general laws, an example of which would be a statement about the thermal conductivity of glass. Now the fact that the phenomenon to be explained can be logically deduced from the general laws with the help of the antecedent conditions constitutes an explanation of that phenomenon. That is, the

⁵Consult Scriven (1959) for alternative approach to explanation and prediction.

⁶Some authorities hold that we never ask "why" in science, but rather "what" and "how."

We do not mean to quibble about these words, for our actual positions probably would not differ to any great extent. If you prefer "what" and "how" to "why" please use them.

way in which we may find out that a given phenomenon can be subsumed under a general law is by determining that the former can be deduced (deductively inferred) from the latter. The schema for accomplishing an explanation can be indicated as follows:

Deductive inference: $\left\{ \begin{array}{l} \text{Statement of the general law(s)} \\ \text{Statement of the antecedent conditions} \end{array} \right.$
 \rightarrow Description of the phenomenon to be explained

In this example it can be seen that the phenomenon to be explained (the immediate drop of the mercury level, followed by its swift rise) may be logically deduced according to the above schema. To illustrate further the nature of explanation, we might develop an analogy using the familiar syllogism concerning Socrates. Say that the phenomenon to be explained is Socrates's death. In the syllogism the two kinds of statements that we require for an explanation are offered. First, the antecedent condition is that "Socrates is a man." And second, the general law is that "All men are mortal." From these two statements, of course, we can deductively infer that Socrates is mortal. With this general understanding of the nature of explanation, let

Deductive inference: $\left\{ \begin{array}{l} \text{General law: All men are mortal} \\ \text{Antecedent condition: Socrates is a man} \end{array} \right.$
 \rightarrow Phenomenon to be explained
 (i.e., Why did Socrates die?): Socrates is mortal

us now ask where the procedure enters the work of the experimental psychologist.

Assume that an experimenter wishes to test the hypothesis that the higher the anxiety the better the performance on a relatively simple task. He decides to take as his measure of anxiety the scores that subjects receive on the Manifest Anxiety Scale (Taylor, 1953) where the higher the score, the greater the anxiety.

Say that the experimenter decides to vary anxiety in two ways. He must select two groups of subjects, one group composed of individuals who have considerable anxiety, a second group of those with little anxiety. A relatively simple task is constructed for the subjects to learn. The evidence report states, in effect, that the high-anxiety group performed better than did the low-anxiety group. The evidence report is thus positive, and since it is in accord with the hypothesis, we may say that the hypothesis is confirmed. His experiment is completed, his problem is solved. But is it really? Although this may be said of the limited problem for which the experiment was conducted, there is still a nagging question — why is his hypothesis "true"? How might it be explained? To answer this question, of course, he must refer to a principle that is more general than his hypothesis. Let us say that he appeals

to a principle, from stimulus-response theory, that says that performance is equal to the amount learned times the drive level present. Letting E stand for performance, H (habit strength) for the learning factor, and D for drive, the principle may be stated as $E = H \times D$.

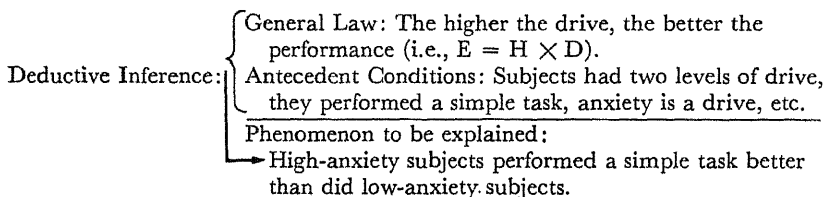
Assuming that anxiety is a specific drive, our experimenter's hypothesis can be established as a specific case of this more general principle. For instance, let us say that the high-anxiety group exhibits a drive factor of 80 units. To simplify matters, assume that both groups learned the task equally well, thus causing the learning factor to be the same for both groups. For instance, H might be 0.50 units. In this event the performance (E) factor for the high-drive group is:

$$E = 0.50 \times 80 = 40.00$$

Assume that the low-anxiety group exhibits a drive factor of 20 units, in which case its performance factor would be:

$$E = 0.50 \times 20 = 10.00$$

Clearly, then, the performance of the high-drive (high-anxiety) group should be superior to the low-drive group, according to this principle. And the principle is quite general in that it ostensibly covers all drives in addition to including a consideration of the learning factor, H . Following our previous schema, then, we have the following situation:



Since it would be possible logically to deduce the experimenter's hypothesis (stated as "the phenomenon to be explained") from the stimulus-response principle together with the necessary antecedent conditions, we may say that the hypothesis is explained.

In Chapter 13 we illustrated the ever-continuing search for a higher-level explanation for general statements. In this chapter we have shown how a relatively specific hypothesis about anxiety and performance can be explained by a more general principle about (1) drives in general and (2) a learning factor (which we ignored because it was not relevant to the present discussion). The next question, obviously, is how to explain this stimulus-response principle. At this point, however, our immediate purpose is accomplished so that we shall conveniently slip off to other topics.

It is important to emphasize that the logical deduction is made on the assumption that the general principle and the statement of the antecedent conditions were actually true. Hence, a more cautious statement about our explanation would be this: Assuming that (1) the general law is true, and (2) the antecedent conditions obtained, then the phenomenon of interest is explained. But how can we be sure that the general principle is, indeed, true? We can never be *absolutely* sure, for it must always assume a probability value. It might someday turn out that the general principle used to explain a particular phenomenon was actually false. In this case what we accepted as a "true" explanation was in reality no explanation at all. Unfortunately, we can do nothing more with this situation — our explanations must always be of a tentative sort. As Feigl has put it, "scientific truths are held only until further notice." We must, therefore, always realize that when we explain a phenomenon it is on the assumption that the general principle used in the explanation is true. If the probability of the general principle is high, then we can feel rather safe. We can, however, never feel absolutely secure. This is merely another indication that we have been given a "probabilistic universe" in which to live. And the sooner we learn to accept this fact (in the present context, the sooner we learn to accept the probabilistic nature of our explanations) the better adjusted to reality we will be.

One final thought on the topic of explanation. We have indicated that an explanation is accomplished by logical deduction. But how frequently do psychologists actually explain their phenomena in such a formal manner? How frequently do they actually cite a general law, state their antecedent conditions, and deductively infer their phenomena from them? The answer, clearly, is that this is done very infrequently. Almost never will you find such a formal process being used in the actual report of scientific investigations. Rather, much more informal methods of reasoning are substituted. One need not set out on his scientific career armed with books of logical formulae and the like. But he should be familiar with the basic logical processes that one could go through in order to accomplish an explanation. As with several matters that we have previously discussed, such as stating hypotheses and evidence reports, it is not necessary that you rigidly follow the procedures that we have set down. What is important, and what we hope you have gained from this discussion, is that you *could* explain a phenomenon in a formal, logical manner if you wanted or needed to.

PREDICTION

The *processes* of making predictions and offering explanations are precisely the same, so that everything we have said about explanation is applicable to prediction. The only difference between explanation and prediction is that a

prediction is made before the phenomenon is observed, whereas explanation occurs after the phenomenon has been recorded. In explanation, then, we start with the phenomenon and logically deduce it from a general law and the attendant antecedent conditions. In prediction, on the other hand, we start with the general law and antecedent conditions and derive our logical consequences. That is, from the general law we infer that a certain phenomenon should occur. We then conduct our experiment to see if it does occur. If it does, then our prediction has been successful. And as we have previously pointed out, this considerably increases the probability of the general law (unless of course the general law already has a very high degree of probability).

To illustrate a possible prediction briefly, say that we are in possession of the general stimulus-response principle that we previously discussed. We might reason thusly: This principle asserts that the higher the drive, the better the performance; anxiety is a specific drive; therefore, we would predict that high-anxiety individuals would perform a given task better than low-anxiety individuals. The conduct of such an experiment would then inform us of the success (or lack of success) of our prediction. Actually, this is precisely what has been done (cf. Spence, Farber, and McFann, 1956).

MISCELLANY

We have attempted to develop logically the major aspects of experimentation throughout the preceding chapters. Starting with the nature of the problem and concluding with prediction of behavior, we have attempted to weave into a coherent pattern the problems and procedures that are important to the experimental psychologist. However, it obviously was not feasible to include there *all* matters of importance. Many of these will simply have to await your future study, but in this chapter we will take up several topics that did not conveniently fit into our general plan of development, though they deserve emphasis.

CONCERNING ACCURACY OF THE DATA ANALYSIS

In one sense, we would like to place this particular section at the beginning of the book, in the boldest type possible. For no matter how much care you give to the other aspects of experimentation, if you are not accurate in your

records and statistical analysis, the experiment is worthless. Unfortunately, there are no set rules that anybody can give you to guarantee accuracy. The best that we can do is to offer you some suggestions which, if followed, will reduce the number of errors, and if you are sufficiently vigilant, eliminate them completely.

The first important point concerns "attitude." Students frequently feel that they must record their data and run their statistical analysis only once, and in so doing, they have amazing confidence in the accuracy of their results. Checking is not for them! Although it is very nice to believe in one's own perfection, the author has observed a large enough number of students and scientists over a sufficiently long period of time to know that this is just not reasonable behavior. We all make mistakes.

The best attitude for a scientist to take is not that he *might* make a mistake, but that he *will* make a mistake;¹ his only problem is where to find it. Accept this suggestion or not, as you like. But remember this: At least the first few times that you run an analysis, the odds are about 99 to 1 that you will make an error. As you become more experienced, the odds might drop to about 10 to 1. The author once had occasion to talk with one of our most outstanding statisticians. To decide a matter it became necessary to run a simple statistical test. Our answer was obviously absurd, so we tried to discover the error. After several checks, however, the fault remained obscure. Finally, a third person, who could look at the problem from a fresh point of view, checked our computations and found the error. The statistician admitted that he was never very good in arithmetic and that he frequently made errors in addition and subtraction.

The first place that an error can be made occurs when you first start to obtain your data. More often than not the experimenter observes behavior and records it by writing it down, so let us take such a case as an example.

Suppose that you are running rats in a T-maze and that you are recording (1) their latency, (2) their running time, and (3) whether they turned left or right. You might take a large piece of paper on which you can identify your subject, and have three columns for your three kinds of data, noting the data for each subject in the appropriate column. Once you indicate the time values and the direction the rat turned, you move on to your next subject; the event is over and there is no possibility for further checking. Hence, any error you make in writing down your data is uncorrectable. You should therefore be exceptionally careful in recording the correct value. You might

¹Unfortunately, since the publication of the first edition of this book there has been empirical documentation of this point for *articles already published in professional journals*. Wolins (1962) reanalyzed data obtained from authors who had published their work, and found several different kinds of errors, including the miscalculation of *F* tests. For example, one author had reported *F* values to be significant when in reality, according to Wolins, they were near one.

fix the value firmly in mind, and then write it down, asking yourself all the time whether you are transcribing the right value. After it is written down, check yourself again to make sure that it is correct. If you find a value that seems particularly out of line, you might double-check it to see if it is right. After double-checking such an unusual datum, it is worthwhile to make a note that it is correct, for later on you might return to it with considerable doubt. For instance, if most of your rats take about two seconds to run the maze, and you write down that one rat had a running time of 57 seconds, take an extra look at the timer to make sure that this reading is correct. Then, if it is, make a little note beside "57 seconds," indicating that the value has been checked.

Frequently, experimenters transcribe the original records of behavior onto another sheet for their statistical analysis. Such a job is long and tedious, and therefore conducive to errors. In recopying data onto new sheets, considerable vigilance must be exercised. The finished job should be checked to make sure that no errors in transcription have been committed. (It is frequently possible to avoid this step. For instance, if you can plan your data sheet so that you can record the measures of behavior directly on the sheet that you will use for your statistical analysis, you will avoid errors of transcription.)

In writing data on a sheet, legibility is of utmost importance, for the reading of numbers is a frequent source of error. You may be surprised at the difficulty you might have in reading your own writing, particularly after a period of time. If you use a pencil, that pencil should be quite sharp and hard, to reduce smudging. If possible, record your data in ink, and if you have to change a number, first make sure it is thoroughly erased or eradicated with ink eradicator.

Labeling of all aspects of your data sheet should be complete, since you may wish to refer to the data at some later time. You should label the experiment clearly, giving its title, the date, place of conduct, and so on. You should unambiguously label each source of data. Your three columns might be labeled, for example, "latency of response in leaving start box," "time in running from start box to close of goal box door," and "direction of turn." Each statistical operation should be clearly labeled. If you run a *t*-test, for instance, the top of your work sheet should state that it is a *t*-test between such and such conditions, using such and such a measure as the dependent variable. In short, label everything pertinent to the records and analysis so that you can return to your work years later and understand it readily.

The actual conduct of the statistical analysis is probably going to be the greatest source of error. It is thus advisable to check each step as you move along. For example, you will probably begin by computing the sums and sums of squares for your groups. Before you substitute these values into your equation you should check them. Otherwise, if they are in error, all of your

later work will have to be redone. Similarly, each multiplication, division, subtraction, and addition should be checked just after it has been made, before you move on to the next operation that incorporates the result. After you have computed your statistical test, checking each step along the way, you should put it aside and do the entire example again, without looking at your previous work. If your second computation, performed independently of the first, checks with your first computation, the probability that you have erred is decreased (it is not eliminated, of course, for you may have made the error twice).

It is advantageous to have someone else conduct the same statistical analysis so that your results and his can be compared. It is also advisable to indicate when you have checked a number or operation. One way to accomplish this is to place a small dot above and to the right of the value (do not place it so low that the dot might be confused with a decimal point). The values of indicating a checked result are: (1) that you can better keep track of where you are in your work, and (2) that at some later time you will know whether or not the work has been checked. Concerning the statistical analysis, another source of errors deserves particular comment. When you are conducting your statistical analysis, some people can easily leave out steps, thus progressing faster. For instance, if your equation calls for you to square a term and then divide that term by the number of subjects, you might tend to do both of these operations at once, merely writing down the result. If you will try *not* to do this, not only will you find that your errors are reduced, but you will be able to check each step of your work more closely. In the above example, for instance, you should write down the square of the number and its divisor. Then write down the result of the division.

RETAIN YOUR ORIGINAL DATA

Your research data are, as you should now be well aware, obtained through the expenditure of a considerable amount of energy and time. They are, therefore, valuable and should be preserved. Furthermore, once the experiment has been published, the data are public and should be made available on request for further advancement of science. Three good reasons for retaining data (with interesting illustrations of each reason) have been suggested by Johnson: "(a) it may be desirable to reanalyze the data from a different point of view; (b) it may be possible to compare data with additional data collected on the same subjects at a later point in time; and (c) it may prove feasible to loan the data to other investigators for their research use" (1964, p. 350).

The importance of this point needs emphasis because of the frequency with which data are not retained, even by professional psychologists. It is amazing to note, for example, that Wolins (1962) contacted 37 authors whose articles appeared in journals of the American Psychological Association. Of 32 who replied, 21 reported that their data had been misplaced, lost, or inadvertently destroyed.

Finally, when you plan to retain your data recall a point made in the previous section. That is, if you look at them years hence, will you be able to understand them? Did you clearly label your data sheets?

COMBINING TESTS OF SIGNIFICANCE

Experimenters frequently have available two or more sets of experimental results that test the same hypothesis. Now, since the experiments are independent of each other, it is possible to combine the results of the statistical tests. Although this procedure may be used for several reasons, one particularly advantageous one is that in neither of the separate experiments was it possible to reject the null hypothesis. Yet the means of the two groups might have been in the same direction, and their differences sufficiently great so that they were strongly suggestive. In such cases it is possible to combine the tests of significance in order to obtain a sounder test of the empirical hypothesis.

A number of techniques for combining two or more tests of significance are available (Lindquist, 1953; Mosteller and Bush, 1954, Chapter 8). Although we cannot possibly go into the advantages and disadvantages of each technique, one approach is extremely easy to use, although it is applicable only to the case where there are but two experiments. To illustrate, say that in one experiment the probability that the null hypothesis is false is 0.07, whereas in the second experiment it is 0.10. Clearly, in neither experiment was it possible to reject the null hypothesis, assuming the significance level was set at 0.05. On the further assumption that the means of the two groups in both experiments were in the same direction (e.g., the experimental group had the higher mean in both cases), we can combine these findings by referring to Figure 15.1. Thus, we locate 0.07 on the horizontal axis, and .10 on the vertical axis (we could, of course, reverse these if we wished). Reading up and across until the lines for the two values of P intersect, we obtain the combined probability. In this example it is less than 0.05, so that, considering the two experiments as combined, we are able to reject the null hypothesis, whereas considering them separately this was not possible.

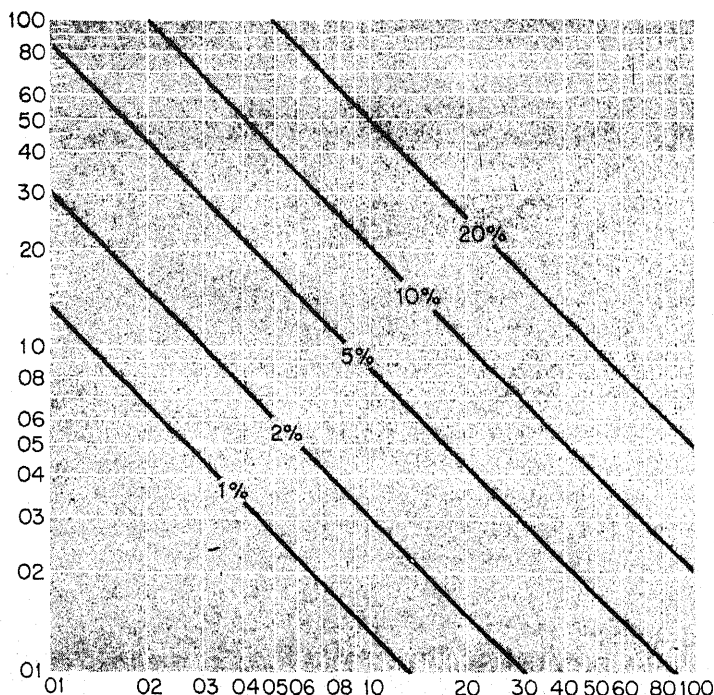


FIGURE 15.1.

The probability of two combined tests of significance. Find the value of one probability along a horizontal axis and the second probability along a vertical axis. Where the vertical line crosses the horizontal line one finds a diagonal that indicates the combined probability value. Five of the more commonly used significance levels are indicated here (from P.C. Baker, personal communication).

ASSUMPTIONS UNDERLYING THE USE OF STATISTICAL TESTS

In applying the statistical tests to the experimental designs presented in this book, one must make certain assumptions. In general, these are that: (1) the dependent variable scores are independent; (2) the variances of the groups are equal (homogeneous); (3) the population distribution is normal; (4) the treatment effects and the error effects are additive. You should be able to get a rough idea of the nature of assumptions 2 and 3. To help you visualize the character of assumption 4, assume that any given dependent variable is a function of two classes of variables — your independent variable and the various extraneous variables. Now we may assume that the dependent variable values due to these two sources of variation can

be expressed as an algebraic sum of the effect of one and the effect of the other, i.e., if R is the response measure used as the dependent variable, if I is the effect of the independent variable, and if E is the combined effect of all of the extraneous variables, then the additivity assumption says that $R = I + E$.

Various tests are available to determine whether or not your particular data allow you to regard assumptions 2, 3, and 4 as tenable, and therefore justify your statistical tests. It does not seem feasible at the present level of approach, however, to elaborate these assumptions nor the nature of the tests for determining whether or not they are satisfied. Certainly you will consider these matters further in your courses in statistics. In addition it is difficult to determine whether or not the assumptions are exactly satisfied, i.e., the tests used for this purpose are rather insensitive. And the consensus is that rather sizeable departures from them can be tolerated still yielding valid statistical analysis (Lindquist, 1953; McNemar, 1962; Dixon and Massey, 1951; Cochran and Cox, 1957; Anderson and Bancroft, 1951; Boneau, 1960, etc.). For further information, you should consult these or a number of other references that are easily available in the area of statistics.

*The first assumption, however, is essential, since each dependent variable score must be independent of every other dependent variable score.*² For example, if one score is 15, the determination that a second score is, say, 10 must in no way be influenced by, or related to, the fact that the first score is 15. If subjects have been selected at random, and if one and only one score on each dependent variable is used for each subject, then the assumption of independence should be satisfied. Frequent violation by students of this assumption is as follows: Suppose that in a learning experiment we conduct, a certain subject yields scores of 10, 8, 6, 5, and 4 on each of five trials. Some students might want to refer to each of these scores as values of X in computing ΣX . Accordingly, they might say that $n = 5$. Now, this is a clear violation of the assumption of independence, for the second score is related to the first score, the third to the first and second, and so on. That is, all of these scores were made by the same subject, and if he happens to be capable, all the scores will tend to be high. A second error that would be committed in this procedure is the artificial inflation of n . That is, these scores are all for one subject; therefore, n cannot possibly equal 5. The proper procedure would be to obtain one single score for this subject, say by adding them up and using the sum as the dependent variable score. Hence, the proper values are $\Sigma X = 33$ and $n = 1$.

We have not emphasized statistical assumptions, except for the one con-

²In the case of the matched groups design, the independence assumption takes a slightly different form, that is that the values of D are independent. Hence, a more adequate statement of this assumption would be that the treatment effects and the error are independent. That is, in terms of the symbols used for the fourth assumption, I and E are independent.

cerning independence, primarily because of the introductory nature of this book. As you progress in statistical and experimental work, you will want to learn more about the assumptions of homogeneity of variances and normality of the populations. With a more thorough knowledge of these assumptions, and how you might determine when they are met, you will be in a better position to evaluate possible errors introduced by violating them. And, of course, when really serious violations of these assumptions occur, you will have to effect some remedy. In general, there are three possible remedies. One is to transform your data (e.g., to take the square root or reciprocal of each score) and then to continue the analysis in the usual way (cf. Edwards, 1968). The second would be to use a type of statistical test known as a nonparametric test, which does not require that you meet the assumptions of normality and homogeneity of variances (cf. Siegel, 1956). Nonparametric tests still require the assumption of independence, however. The third is to adjust your level of probability. The logic here is that your P value for your statistical test is not a true one when the assumptions for it have been seriously violated; but it is possible to adjust that value so that it more closely approximates the true probability level (cf. Horsnell, 1953).

REDUCING ERROR VARIANCE

The purpose of this section is to consider ways in which the experimenter can increase his chances of rejecting the null hypothesis, *if in fact it should be rejected*. The point may be illustrated by taking two extremes. First, if an experimenter conducts a "sloppy" experiment (e.g., his controls are poor) his chances of rejecting a null hypothesis that is false are quite low. On the other hand, if he conducts a highly refined (e.g., well controlled) experiment, he increases his chances of rejecting a false null hypothesis. In short, if there is a lot of "noise" in the experimental situation, the dependent variable scores are going to vary for a lot of reasons other than the fact that the independent variable varied, thus obscuring any systematic relationship between the independent and dependent variable. There are two general ways in which an experimenter can increase his chances of rejecting a null hypothesis that really should be rejected. To understand them, let us get the basic equation for the t -test for the two-randomized-groups design before us:³

$$(15.1) \quad t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

³See footnote 12, p. 180.

Now, we know that the larger the value of t , the greater the likelihood that we will be able to reject the null hypothesis. Hence, our question is: How can we design an experiment such that the value of t can legitimately be increased? In other words, how can we increase the numerator of Equation (15.1) and decrease the denominator? We have previously seen that the numerator can often be increased by exaggerating the difference in the independent variable values that you administer to your two groups. For instance, if we are seeking to determine whether amount of practice affects amount learned, we are more likely to obtain a significant difference between two groups if we have them practice 100 trials versus 10 trials, than if we have them practice 15 trials versus 10 trials. This is so because we would probably increase the difference between the means on the dependent variable of the two groups and, as we said, the greater the mean difference, the larger the value of t . Let us now consider the denominator of Equation (15.1).

In every experiment there is a certain error variance, and in our statistical analysis we obtain an estimate of it. In the two-groups designs, the error variance is the denominator of the t ratio. Where we use Duncan's range test, our estimate of the error variance is provided by Equation (9.5) (p. 212), which is an equation for computing the square root of the error variance. And in factorial designs, the error variance is the denominator of the F ratio. Basically, these three estimates of the error variance are measures of the same thing — the extent to which subjects treated alike exhibit variability in their dependent variable scores. There are many reasons why we obtain different scores for subjects treated alike. For one, subjects are all "made" differently, and they will all react somewhat differently to the same experimental treatment. For another, it simply is impossible to treat all subjects in the same group precisely the same; we always have a number of extraneous variables operating on our subjects in a random manner, resulting in differential effects. And finally, some of the error variance is due to imperfections in our measuring devices. No device can provide a completely "true" score, nor can we as humans make completely accurate and consistent readings of the measuring device.

In many ways, it is unfortunate that dependent variable scores for subjects treated alike show so much variability, but we must learn to live with the fact that this variability is with us. The best we can do is attempt to reduce it. Let us emphasize the reason that we want to reduce the error variance in an experiment. The illustration will be in terms of the t -test as used in a two-groups design, but what we have to say may be considered applicable to other designs and their statistical tests. Let us say that we know that the difference between the means of two groups is five. Consider two situations, one where the error variance is rather large, and one where it is rather small. For example, say that the error variance is five in the first case, but two in the second. For the first case, then, our computed t would be: $t = \frac{5}{\sqrt{5}} = 1.0$,

and for the second it would be $t = \frac{5}{2} = 2.5$. Clearly, in the first case we would *fail* to reject our null hypothesis, while in the second case we are likely to reject it. For both situations we have the same mean difference. Such a difference in our two experiments is, of course, of the utmost importance. We naturally seek to reject the null hypothesis, if it "should" be rejected. What we mean is that the null hypothesis specifies that there is no difference between our two groups. However, in a given experiment we find that the means of our groups differ. The question is whether or not the difference is sufficiently large to allow us to reject the null hypothesis. Now if our error variance is sufficiently large, we shall not. But it may be rejected if the error variance is sufficiently small. Hence we seek to obtain an error variance that is sufficiently small to allow us to reject our null hypothesis. If after reducing it as much as possible, we still cannot reject our null hypothesis, then it seems reasonable to conclude that the null hypothesis should actually not be rejected.

This point emphasizes that we are *not* trying to find out how to increase our chances of rejecting the null hypothesis in a biased sort of way; we only want to increase our chances of rejecting the null hypothesis if it really should be rejected. Let us now consider ways in which we can reduce the error variance in our experiments. To do this we shall consider the denominator of Equation (15.1) in greater detail. First, we can see clearly that as the variances (i.e., s_1^2 and s_2^2) of the groups decrease, the size of t increases. To illustrate, assume that the mean difference is 5, that the variances are each 64, and that n_1 and n_2 are both 8. In this event,

$$t = \frac{5}{\sqrt{\frac{64}{8} + \frac{64}{8}}} = 1.25$$

Now let us say that the experiment is conducted again, except that this time we are able to reduce the variances to 16. In this case,

$$t = \frac{5}{\sqrt{\frac{16}{8} + \frac{16}{8}}} = 2.50$$

Granting, then, that it is highly advisable to reduce the variances of our groups, how can we accomplish this? There are several possibilities. First, recall that our subjects, when they enter the experimental situation, are all different, and that the larger such differences, the greater the variances of our groups. Therefore, one obvious way to reduce the variances of our groups, and hence the error variance, is to reduce the extent to which our subjects are different. Psychologists frequently increase the homogeneity of their groups by selection. For example, we work with a number of different strains of rats. In any given experiment, however, all the subjects are usually

taken from a single strain — the Wistar strain, the Sprague-Dawley strain, or whatever. If a psychologist randomly assigns rats from several different strains to his groups, he is probably going to increase his variances. Working with humans is more difficult, but even here the selection of subjects that are similar is a frequent practice, and should be considered. For example, using college students as subjects undoubtedly results in smaller variances than if we selected subjects at random from the general population of humans. But you could even be selective in your college population; you might use only females, only subjects with IQs above 120, only students with low anxiety scores, and so on.

At this point, one serious objection that comes to mind is that by selection of subjects you thus restrict the extent to which you can generalize your results. Thus, if you sample only high-IQ students, you will certainly be in danger if you try to generalize your findings to low-IQ students, or to any other population that you have not sampled. For this reason, selection of homogeneous subjects for only two groups in an experiment (e.g., experimental vs. control groups) should be seriously pondered before it is adopted. For the greater the extent to which you select homogeneous subjects, the less you can generalize.

The solution to this problem could come from a systematic sampling of a population in which a number of different values for the population are incorporated in the experiment. In this case a factorial design is entailed, as in Table 14.2. For instance, you might classify your subjects as high IQ, medium IQ, or low IQ (this is called *stratification*). In this case you have three homogenous groups, and you can also generalize to your subject population as far as IQ is concerned. Furthermore, and this is a somewhat more advanced point, you can reduce your error variance by using this type of design, i.e., when you classify subjects into levels (the vertical cells of the factorial), you compute the variation in the dependent variable due to levels (here three levels of IQ); then this variation is “automatically” taken out of the error term, resulting in an increase in the precision of the experiment. As we said, this is a somewhat more advanced point than is really appropriate here, but perhaps by being alerted to it, your future work will be facilitated. In any event, what we have said is that: (1) error variance can be decreased by selecting homogenous subjects; (2) but if you use only two groups you restrict the extent to which you can generalize; (3) therefore, you can stratify your subjects by repeating your experiment at different levels of homogeneous subjects, thus decreasing your error variance while maintaining the extent to which you can generalize.

A *second* way in which you can reduce your variances is in your experimental procedure. The ideal is to treat all subjects in the same group as precisely alike as possible. We cannot emphasize this too strongly. We have counseled the use of a tape recorder for administering instructions, in order

that all subjects would receive precisely the same words, with precisely the same intonations. If you rather casually tell your subject what to do, varying the way in which you say it with different subjects, you are probably increasing your variances. Similarly, the greater the number of extraneous variables that are operating in a random fashion, the greater will be your variances. If, for example, noises are present in varying degrees for some subjects, but not present at all for others, your group variances are probably going to increase. Here again, however, you should recognize that if you eliminate extraneous variables to any great extent, you will have difficulty in generalizing to situations where they are present. For example, if all your subjects are run in sound-deadened rooms, then you should not, strictly speaking, generalize to situations where noises are randomly present. But since we usually are not trying to generalize, at least not immediately, to such uncontrolled stimulus conditions, this general objection need not greatly disturb us.

A *third* way to reduce your variances concerns the errors that you might make — errors in reading your measuring instruments, in recording your data, and in your statistical analysis. The more errors that are present, the larger will be the variances, assuming that such errors are of a random nature. This point also relates to the matter of the reliability of your dependent variable, or perhaps more appropriately to how reliably you measured it, as discussed on pp. 157–159. Hence, the more reliable your measures of the dependent variable, the less will be your error variance. One way in which the reliability of the dependent variable measure can be increased is to make more than one observation on each subject; if your experimental procedure allows this you would be wise to consider it.

The three techniques noted above are ways of reducing the error variance by reducing the variances of your groups. Another possible technique for reducing the error variance concerns the design that you select. The clearest example for the designs that we have considered would be to replace the two-randomized-groups design with the matched-groups design for two groups, providing that there is a substantial correlation between the independent variable and dependent variable. As you may recall, the error variance is reduced in accordance with the size of the correlation (p. 186). The factorial design can also be used to decrease your error variance. For example, you might incorporate an otherwise extraneous variable in your design and remove the variance attributable to that variable from your error variance. This was the point made above for IQ of subjects (p. 359).

A technique that is frequently effective in reducing error variance is the “analysis of covariance.” Briefly, this technique enables you to obtain a measure of what you think is a particularly relevant extraneous variable that you are not controlling. This usually has to do with some characteristic of your subjects. For instance, if you are conducting a study of the effect of certain psychological variables on weight, you might use as your measure

the weight of your subjects before you administer your experimental treatments. Through analysis of covariance, you then can "statistically control" this variable; that is, you can remove the effect of initial weight from your dependent variable scores, thus decreasing your error variance. We might note that the degree of success in reducing error variance with the analysis of covariance depends on the size of the correlation between your extraneous variable and your dependent variable. In your future study of experimentation and statistics you should attempt to learn how this technique is applied.

We may note that Ray (1960, Chapter 17) offers an excellent discussion of the use of various designs for reducing error variance. Let us summarize his presentation by starting with the notion of *precision* of a design: The more precise a design, the less the error variance resulting from its use. Ray considers five different designs: (1) a randomized-groups design (as in Chapters 5, 9, and 10) which is analyzed by means of an analysis of variance; (2) a gains design in which there is an initial measure of proficiency for each subject and a post training measure for him; the dependent variable measure is then the amount of gain for each subject, which value is obtained by subtracting the post training score from the pretraining score (as discussed on pp. 290-292); (3) an analysis of covariance design; (4) a matching design (Chapter Eight); and (5) a repeated treatments design — in this case the same subject is exposed to several experimental treatments, as in the use of counterbalancing (see pp. 132-135). Ray takes as his standard the randomized-groups design and ascertains the effectiveness (in terms of amount of precision) of each of the other designs relative to that of the randomized-groups design. His conclusions are: (1) all of the latter four designs are more precise than the randomized-groups design; (2) of the latter four designs, the repeated treatments design is generally superior to the gains design, to the covariance design, and to the matching design; and (3) the differences in precision between a matching and a covariance design are not very great. For more specific conclusions, you should refer to Ray; in particular, note that the superiority in precision of some designs depends on the value of the correlation that obtains in your experiment (e.g., between your matching and your dependent variable). The general point here is that error variance *can* be reduced by judicious selection of the design that you use.

Referring back to Equation (15.1) (p. 356) we have seen that as the variances of our groups decrease, the error variance decreases, and the size of t increases. The other factor in the denominator is n . As the size of n increases, the error variance decreases. This occurs for two reasons: because we use n in the computation of variances, and because n is also used otherwise. We might comment that increasing the number of subjects per group is probably one of the easiest ways to decrease, usually very sharply, the error variance. More will be said shortly about the number of subjects in an experiment.

We have now considered a number of ways in which the error variance of an experiment can be reduced. They should all provide useful hints for you as you develop your experimental plan. Some of the methods suggested should be quite easy to incorporate, while others will no doubt be unrealistic for your particular experiment. Furthermore, some of the procedures are generally more effective than others. One of the more interesting and promising developments in statistics and experimental design is the attempt to ascertain the relatively more effective procedures for decreasing error variance. Even though this topic is somewhat more advanced than is appropriate here, perhaps a word or two will sensitize you to the issue for your future learning. Overall and Dalal (1965), for example, present a very valuable analysis of the factors that contribute to error variance. In their terms, the problem is to maximize the power of the experiment; more completely, they attempt to ascertain the relative effects of several variables on increasing the treatment effects relative to the experimental error. Among the variables that they consider are: (1) the number of independent observations made on each subject; (2) the number of subjects per group; (3) the number of levels of a control variable (e.g., the number of values of some variable of secondary interest that is incorporated into a factorial design as in Chapter 10); (4) the number of places that the experiment is repeated (e.g., the same experiment may be conducted in several different hospitals or schools); and (5) the number of treatment conditions in the experiment. These researchers then present tables in which the power of the experiment may be determined as a function of the particular value assumed by such variables. That is, as these kinds of variables vary (e.g., as the number of subjects per group varies) the relative power of the experiment can be ascertained; in this way the experimenter can enter the tables with the particular values that he has planned on using, and determine the relative power of his experiment. He can also see whether certain changes in his experiment (e.g., increasing the number of subjects by a certain amount) will have a noticeable effect on the power. Among the general conclusions offered by Overall and Dalal, some are particularly relevant (if unsurprising) here. For example, the number of observations made on each subject, the number of subjects per group, and the number of locations at which the experiment was conducted all increase the likelihood of rejecting the null hypothesis (if it should be rejected). But, other things being equal, repeating the experiment in different locations has a relatively greater effect in increasing the estimated value of F (in the F -test). Another conclusion concerns the relative value of increasing the number of subjects as against increasing the number of observations on each subject. Their answer is that "... it is unquestionably better to use more S 's and test each one only once ... no matter how unreliable the measurement involved" (Overall and Dalal, 1965, p. 347).

These are but illustrations of the way in which our knowledge in the area of experimental design is increasing. As the results of continued research on

this topic increase, we can well expect that the power of experiments of the future will dwarf those of today.

We have tried in this section to indicate the importance of the reduction of error variance in experimentation and to suggest some of the ways that it might be accomplished. Unfortunately, it is not possible to provide an exhaustive coverage of the available techniques, because of both lack of space and complexities that would take us beyond our present level of discussion. Excellent treatments of this topic have been given elsewhere, although they require a somewhat advanced knowledge of experimentation and statistics (e.g., Cochran and Cox, 1957; Fisher, 1953).

Let us conclude this most important topic by summarizing the more important points that have been made. First, the likelihood of rejecting the null hypothesis can be increased by increasing the difference between the values of the independent variable administered to the groups in the experiment, and by decreasing the error variance. Specific ways that one is likely to decrease his error variance are: (1) classify (stratify) subjects into homogeneous levels according to their scores on some relevant measure; (2) standardize, in a strict fashion, the experimental procedures used; (3) reduce errors in observing and recording the dependent variable scores (and make more than one measurement on each subject if practicable); (4) select a relatively precise design; (5) increase the number of subjects per group; (6) replicate the experiment.

NUMBER OF SUBJECTS PER GROUP

"How many subjects should I have in my groups?" is a question that students usually ask in a beginning course in experimental psychology. We have touched on this question above, but there are several additional points that can profitably be made. First let us note a rather traditional procedure often used by experimenters. That is to run a number of subjects, more or less arbitrarily determined, and see how the results turn out. If the groups differ significantly, the experimenter may be satisfied with the numbers he chose or additional subjects may be run to confirm the significant findings. On the other hand, if the groups do not differ significantly, but the differences are promising, more subjects may be run in the hope that the additional data will produce significance.⁴

⁴This latter procedure cannot be legitimately (strictly speaking) defended in other than a preliminary investigation. Clearly, one who keeps running subjects until he obtains a significant difference may capitalize on chance. For example, if one runs ten subjects per group and obtains a t value that approaches significance, he might run ten more subjects per group. Assume that the t is now significant. But the results of these additional subjects might be due merely to chance. The experiment is stopped and success proclaimed. If still more subjects were run, however, significance would be lost, and the experimenter would never know this fact. If such an experiment is to be cross-validated, this procedure is, of course, legitimate.

Although we cannot adequately answer the student's question, we can offer some guiding considerations. First, we have seen that the larger the number of subjects run, the more reliably can be estimated the difference (if such exists) between the groups. This is a true and sure statement, but it does not help very much. We can clearly say that 100 subjects per group is better than 50. You may want to know if 20 subjects per group is enough. That depends, first of all, on the "true" (population) difference between your groups, and second, on the size of the variances of your groups. What we can say is that the larger the true difference between groups, the smaller the number of subjects required for the experiment; and the smaller the group variances, the fewer subjects required. Now if you know what the differences are, and also what the variances are, the number of subjects required can be estimated. Unfortunately, experimenters do not usually have this information, or if they have it they do not consider the matter worth the effort required to answer the question. We shall not attempt to judge what should or should not be done in this respect but shall illustrate the procedure for determining the minimum number of subjects required, given these two bits of information. (Possible sources of this information include: (1) an experiment reported in the literature similar to the one you want to run, from which you can abstract the necessary information, or better, (2) a pilot study conducted by yourself to yield estimates of the information that is needed.)

In any event, let us suppose that you are going to conduct a two-randomized-groups experiment. You estimate (on the basis of previously collected data) that the mean score of Condition A is 10, and that the mean of Condition B is 15. The difference between these means is 5. You also estimate that the variances of your two groups are both 75. Say that you set your significance level at .05, in which case the value of t that you will need to reject the null hypothesis is *approximately* 2 (you may be more precise if you like). Assume that you want an equal number of subjects in both groups. Now we have this information:

$$\begin{aligned}\bar{X}_1 - \bar{X}_2 &= 5 \\ s_1^2 \text{ and } s_2^2 \text{ both} &= 75 \\ t &= 2\end{aligned}$$

Let us solve Equation (15.1) for n instead of for t . By simple algebraic manipulation we find that, on the above assumptions, Equation (15.1) becomes:

$$(15.2) \quad n = \frac{2t^2s^2}{(\bar{X}_1 - \bar{X}_2)^2}$$

Substituting the above values in Equation (15.2) and solving for n , we find:

$$n = \frac{2(2)^2(75)}{(15 - 10)^2} = \frac{600}{25} = 24$$

We can say, therefore, that with this true mean difference, and with these variances for our two groups, and using the 5 per cent level of significance, we need a minimum of 24 subjects per group to achieve significance. We have only approximated the value of t necessary at the 5 per cent level, however, and we have not allowed for any possible increase in the variance of our two groups. Therefore, we should underline the word *minimum*. To be safe, then, we should probably run somewhat more than 24 subjects per group; 30 would seem reasonable in this case.⁵

We have illustrated a technique for estimating the minimum number of subjects necessary to reject the null hypothesis, if in fact you should reject it, for the case of the two-randomized-groups design. The technique can also be extended, with a little industry, to the equations for analyzing the other designs that we have taken up.

A LOOK TO THE FUTURE

This concludes our presentation. You have finished a book on Experimental Psychology, but the topic itself is endless. Among those who have studied this book, some will go on to become talented researchers; we hope that those who do will themselves discover some new and interesting characteristics of behavior. For all, we hope that an increased appreciation for sound psychological knowledge was gained.

⁵This procedure is offered only as a rough guide, for we are neglecting power considerations of the statistical test. This procedure has a minimal power for rejecting the null hypothesis.

SQUARES AND SQUARE ROOTS

N	N ²	√N	√10N
1.00	1.0000	1.00000	3.16228
1.01	1.0201	1.00499	3.17805
1.02	1.0404	1.00995	3.19374
1.03	1.0609	1.01489	3.20936
1.04	1.0816	1.01980	3.22490
1.05	1.1025	1.02470	3.24037
1.06	1.1236	1.02956	3.25576
1.07	1.1449	1.03441	3.27109
1.08	1.1664	1.03923	3.28634
1.09	1.1881	1.04403	3.30151
1.10	1.2100	1.04881	3.31662
1.11	1.2321	1.05357	3.33167
1.12	1.2544	1.05830	3.34664
1.13	1.2769	1.06301	3.36155
1.14	1.2996	1.06771	3.37639
1.15	1.3225	1.07238	3.39116
1.16	1.3456	1.07703	3.40588
1.17	1.3689	1.08167	3.42053
1.18	1.3924	1.08628	3.43511
1.19	1.4161	1.09087	3.44964
1.20	1.4400	1.09545	3.46410
1.21	1.4641	1.10000	3.47851
1.22	1.4884	1.10454	3.49285
1.23	1.5129	1.10905	3.50714
1.24	1.5376	1.11355	3.52136
1.25	1.5625	1.11803	3.53553
1.26	1.5876	1.12250	3.54965
1.27	1.6129	1.12694	3.56371
1.28	1.6384	1.13137	3.57771
1.29	1.6641	1.13578	3.59166
1.30	1.6900	1.14018	3.60555
1.31	1.7161	1.14455	3.61939
1.32	1.7424	1.14891	3.63318
1.33	1.7689	1.15326	3.64692
1.34	1.7956	1.15758	3.66060
1.35	1.8225	1.16190	3.67423
1.36	1.8496	1.16619	3.68782
1.37	1.8769	1.17047	3.70135
1.38	1.9044	1.17473	3.71484
1.39	1.9321	1.17898	3.72827
1.40	1.9600	1.18322	3.74166
1.41	1.9881	1.18743	3.75500
1.42	2.0164	1.19164	3.76829
1.43	2.0449	1.19583	3.78153
1.44	2.0736	1.20000	3.79473
1.45	2.1025	1.20416	3.80789
1.46	2.1316	1.20830	3.82099
1.47	2.1609	1.21244	3.83406
1.48	2.1904	1.21655	3.84708
1.49	2.2201	1.22066	3.86005
1.50	2.2500	1.22474	3.87298
N	N ²	√N	√10N

N	N ²	√N	√10N
1.50	2.2500	1.22474	3.87298
1.51	2.2801	1.22882	3.88587
1.52	2.3104	1.23288	3.89872
1.53	2.3409	1.23693	3.91152
1.54	2.3716	1.24097	3.92428
1.55	2.4025	1.24499	3.93700
1.56	2.4336	1.24900	3.94968
1.57	2.4649	1.25300	3.96232
1.58	2.4964	1.25698	3.97492
1.59	2.5281	1.26095	3.98748
1.60	2.5600	1.26491	4.00000
1.61	2.5921	1.26886	4.01248
1.62	2.6244	1.27279	4.02492
1.63	2.6569	1.27671	4.03733
1.64	2.6896	1.28062	4.04969
1.65	2.7225	1.28452	4.06202
1.66	2.7556	1.28841	4.07431
1.67	2.7889	1.29228	4.08656
1.68	2.8224	1.29615	4.09878
1.69	2.8561	1.30000	4.11096
1.70	2.8900	1.30384	4.12311
1.71	2.9241	1.30767	4.13521
1.72	2.9584	1.31149	4.14729
1.73	2.9929	1.31529	4.15933
1.74	3.0276	1.31909	4.17133
1.75	3.0625	1.32288	4.18330
1.76	3.0976	1.32665	4.19524
1.77	3.1329	1.33041	4.20714
1.78	3.1684	1.33417	4.21900
1.79	3.2041	1.33791	4.23084
1.80	3.2400	1.34164	4.24264
1.81	3.2761	1.34536	4.25441
1.82	3.3124	1.34907	4.26615
1.83	3.3489	1.35277	4.27785
1.84	3.3856	1.35647	4.28952
1.85	3.4225	1.36015	4.30116
1.86	3.4596	1.36382	4.31277
1.87	3.4969	1.36748	4.32435
1.88	3.5344	1.37113	4.33590
1.89	3.5721	1.37477	4.34741
1.90	3.6100	1.37840	4.35890
1.91	3.6481	1.38203	4.37035
1.92	3.6864	1.38564	4.38178
1.93	3.7249	1.38924	4.39318
1.94	3.7636	1.39284	4.40454
1.95	3.8025	1.39642	4.41588
1.96	3.8416	1.40000	4.42719
1.97	3.8809	1.40357	4.43847
1.98	3.9204	1.40712	4.44972
1.99	3.9601	1.41067	4.46094
2.00	4.0000	1.41421	4.47214
N	N ²	√N	√10N

SQUARES AND SQUARE ROOTS—(Continued)

N	N ²	\sqrt{N}	$\sqrt{10N}$
2.00	4.0000	1.41421	4.47214
2.01	4.0401	1.41774	4.48330
2.02	4.0804	1.42127	4.49444
2.03	4.1209	1.42478	4.50555
2.04	4.1616	1.42829	4.51664
2.05	4.2025	1.43178	4.52769
2.06	4.2436	1.43527	4.53872
2.07	4.2849	1.43875	4.54973
2.08	4.3264	1.44222	4.56070
2.09	4.3681	1.44568	4.57165
2.10	4.4100	1.44914	4.58258
2.11	4.4521	1.45258	4.59347
2.12	4.4944	1.45602	4.60435
2.13	4.5369	1.45945	4.61519
2.14	4.5796	1.46287	4.62601
2.15	4.6225	1.46629	4.63681
2.16	4.6656	1.46969	4.64758
2.17	4.7089	1.47309	4.65833
2.18	4.7524	1.47648	4.66905
2.19	4.7961	1.47986	4.67974
2.20	4.8400	1.48324	4.69042
2.21	4.8841	1.48661	4.70106
2.22	4.9284	1.48997	4.71169
2.23	4.9729	1.49332	4.72229
2.24	5.0176	1.49666	4.73286
2.25	5.0625	1.50000	4.74342
2.26	5.1076	1.50333	4.75395
2.27	5.1529	1.50665	4.76445
2.28	5.1984	1.50997	4.77493
2.29	5.2441	1.51327	4.78539
2.30	5.2900	1.51658	4.79583
2.31	5.3361	1.51987	4.80625
2.32	5.3824	1.52315	4.81664
2.33	5.4289	1.52643	4.82701
2.34	5.4756	1.52971	4.83735
2.35	5.5225	1.53297	4.84768
2.36	5.5696	1.53623	4.85798
2.37	5.6169	1.53948	4.86826
2.38	5.6644	1.54272	4.87852
2.39	5.7121	1.54596	4.88876
2.40	5.7600	1.54919	4.89898
2.41	5.8081	1.55242	4.90918
2.42	5.8564	1.55563	4.91935
2.43	5.9049	1.55885	4.92950
2.44	5.9536	1.56205	4.93964
2.45	6.0025	1.56525	4.94975
2.46	6.0516	1.56844	4.95984
2.47	6.1009	1.57162	4.96991
2.48	6.1504	1.57480	4.97996
2.49	6.2001	1.57797	4.98999
2.50	6.2500	1.58114	5.00000
N	N ²	\sqrt{N}	$\sqrt{10N}$

N	N ²	\sqrt{N}	$\sqrt{10N}$
2.50	6.2500	1.58114	5.00000
2.51	6.3001	1.58430	5.00999
2.52	6.3504	1.58745	5.01996
2.53	6.4009	1.59060	5.02991
2.54	6.4516	1.59374	5.03984
2.55	6.5025	1.59687	5.04975
2.56	6.5536	1.60000	5.05964
2.57	6.6049	1.60312	5.06952
2.58	6.6564	1.60624	5.07937
2.59	6.7081	1.60935	5.08920
2.60	6.7600	1.61245	5.09902
2.61	6.8121	1.61555	5.10882
2.62	6.8644	1.61864	5.11859
2.63	6.9169	1.62173	5.12835
2.64	6.9696	1.62481	5.13809
2.65	7.0225	1.62788	5.14782
2.66	7.0756	1.63095	5.15752
2.67	7.1289	1.63401	5.16720
2.68	7.1824	1.63707	5.17687
2.69	7.2361	1.64012	5.18652
2.70	7.2900	1.64317	5.19615
2.71	7.3441	1.64621	5.20577
2.72	7.3984	1.64924	5.21536
2.73	7.4529	1.65227	5.22494
2.74	7.5076	1.65529	5.23450
2.75	7.5625	1.65831	5.24404
2.76	7.6176	1.66132	5.25357
2.77	7.6729	1.66433	5.26308
2.78	7.7284	1.66733	5.27257
2.79	7.7841	1.67033	5.28205
2.80	7.8400	1.67332	5.29150
2.81	7.8961	1.67631	5.30094
2.82	7.9524	1.67929	5.31037
2.83	8.0089	1.68226	5.31977
2.84	8.0656	1.68523	5.32917
2.85	8.1225	1.68819	5.33854
2.86	8.1796	1.69115	5.34790
2.87	8.2369	1.69411	5.35724
2.88	8.2944	1.69706	5.36656
2.89	8.3521	1.70000	5.37587
2.90	8.4100	1.70294	5.38516
2.91	8.4681	1.70587	5.39444
2.92	8.5264	1.70880	5.40370
2.93	8.5849	1.71172	5.41295
2.94	8.6436	1.71464	5.42218
2.95	8.7025	1.71756	5.43139
2.96	8.7616	1.72047	5.44059
2.97	8.8209	1.72337	5.44977
2.98	8.8804	1.72627	5.45894
2.99	8.9401	1.72916	5.46809
3.00	9.0000	1.73205	5.47723
N	N ²	\sqrt{N}	$\sqrt{10N}$

SQUARES AND SQUARE ROOTS—(Continued)

N	N ²	\sqrt{N}	$\sqrt{10N}$
3.00	9.0000	1.73205	5.47723
3.01	9.0601	1.73494	5.48635
3.02	9.1204	1.73781	5.49545
3.03	9.1809	1.74069	5.50454
3.04	9.2416	1.74356	5.51362
3.05	9.3025	1.74642	5.52268
3.06	9.3636	1.74929	5.53173
3.07	9.4249	1.75214	5.54076
3.08	9.4864	1.75499	5.54977
3.09	9.5481	1.75784	5.55878
3.10	9.6100	1.76068	5.56776
3.11	9.6721	1.76352	5.57674
3.12	9.7344	1.76635	5.58570
3.13	9.7969	1.76918	5.59464
3.14	9.8596	1.77200	5.60357
3.15	9.9225	1.77482	5.61249
3.16	9.9856	1.77764	5.62139
3.17	10.0489	1.78045	5.63028
3.18	10.1124	1.78326	5.63915
3.19	10.1761	1.78606	5.64801
3.20	10.2400	1.78885	5.65685
3.21	10.3041	1.79165	5.66569
3.22	10.3684	1.79444	5.67450
3.23	10.4329	1.79722	5.68331
3.24	10.4976	1.80000	5.69210
3.25	10.5625	1.80278	5.70088
3.26	10.6276	1.80555	5.70964
3.27	10.6929	1.80831	5.71839
3.28	10.7584	1.81108	5.72713
3.29	10.8241	1.81384	5.73585
3.30	10.8900	1.81659	5.74456
3.31	10.9561	1.81934	5.75326
3.32	11.0224	1.82209	5.76194
3.33	11.0889	1.82483	5.77062
3.34	11.1556	1.82757	5.77927
3.35	11.2225	1.83030	5.78792
3.36	11.2896	1.83303	5.79655
3.37	11.3569	1.83576	5.80517
3.38	11.4244	1.83848	5.81378
3.39	11.4921	1.84120	5.82237
3.40	11.5600	1.84391	5.83095
3.41	11.6281	1.84662	5.83952
3.42	11.6964	1.84932	5.84808
3.43	11.7649	1.85203	5.85662
3.44	11.8336	1.85472	5.86515
3.45	11.9025	1.85742	5.87367
3.46	11.9716	1.86011	5.88218
3.47	12.0409	1.86279	5.89067
3.48	12.1104	1.86548	5.89915
3.49	12.1801	1.86815	5.90762
3.50	12.2500	1.87083	5.91608
N	N ²	\sqrt{N}	$\sqrt{10N}$

N	N ²	\sqrt{N}	$\sqrt{10N}$
3.50	12.2500	1.87083	5.91608
3.51	12.3201	1.87350	5.92463
3.52	12.3904	1.87617	5.93296
3.53	12.4609	1.87883	5.94138
3.54	12.5316	1.88149	5.94979
3.55	12.6025	1.88414	5.95819
3.56	12.6736	1.88680	5.96657
3.57	12.7449	1.88944	5.97495
3.58	12.8164	1.89209	5.98331
3.59	12.8881	1.89473	5.99166
3.60	12.9600	1.89737	6.00000
3.61	13.0321	1.90000	6.00833
3.62	13.1044	1.90263	6.01664
3.63	13.1769	1.90526	6.02495
3.64	13.2496	1.90788	6.03324
3.65	13.3225	1.91050	6.04152
3.66	13.3956	1.91311	6.04979
3.67	13.4689	1.91572	6.05805
3.68	13.5424	1.91833	6.06630
3.69	13.6161	1.92094	6.07454
3.70	13.6900	1.92354	6.08276
3.71	13.7641	1.92614	6.09098
3.72	13.8384	1.92873	6.09918
3.73	13.9129	1.93132	6.10737
3.74	13.9876	1.93391	6.11555
3.75	14.0625	1.93649	6.12372
3.76	14.1376	1.93907	6.13188
3.77	14.2129	1.94165	6.14003
3.78	14.2884	1.94422	6.14817
3.79	14.3641	1.94679	6.15630
3.80	14.4400	1.94936	6.16441
3.81	14.5161	1.95192	6.17252
3.82	14.5924	1.95448	6.18061
3.83	14.6689	1.95704	6.18870
3.84	14.7456	1.95959	6.19677
3.85	14.8225	1.96214	6.20484
3.86	14.8996	1.96469	6.21289
3.87	14.9769	1.96723	6.22093
3.88	15.0544	1.96977	6.22896
3.89	15.1321	1.97231	6.23699
3.90	15.2100	1.97484	6.24500
3.91	15.2881	1.97737	6.25300
3.92	15.3664	1.97990	6.26099
3.93	15.4449	1.98242	6.26897
3.94	15.5236	1.98494	6.27694
3.95	15.6025	1.98746	6.28490
3.96	15.6816	1.98997	6.29285
3.97	15.7609	1.99249	6.30079
3.98	15.8404	1.99499	6.30872
3.99	15.9201	1.99750	6.31664
4.00	16.0000	2.00000	6.32456
N	N ²	\sqrt{N}	$\sqrt{10N}$

SQUARES AND SQUARE ROOTS—(Continued)

N	N ²	√N	√10N
4.00	16.0000	2.00000	6.32456
4.01	16.0801	2.00250	6.33246
4.02	16.1604	2.00499	6.34035
4.03	16.2409	2.00749	6.34823
4.04	16.3216	2.00998	6.35610
4.05	16.4025	2.01246	6.36396
4.06	16.4836	2.01494	6.37181
4.07	16.5649	2.01742	6.37966
4.08	16.6464	2.01990	6.38749
4.09	16.7281	2.02237	6.39531
4.10	16.8100	2.02485	6.40312
4.11	16.8921	2.02731	6.41093
4.12	16.9744	2.02978	6.41872
4.13	17.0569	2.03224	6.42651
4.14	17.1396	2.03470	6.43428
4.15	17.2225	2.03715	6.44205
4.16	17.3056	2.03961	6.44981
4.17	17.3889	2.04206	6.45755
4.18	17.4724	2.04450	6.46529
4.19	17.5561	2.04695	6.47302
4.20	17.6400	2.04939	6.48074
4.21	17.7241	2.05183	6.48845
4.22	17.8084	2.05426	6.49615
4.23	17.8929	2.05670	6.50384
4.24	17.9776	2.05913	6.51153
4.25	18.0625	2.06155	6.51920
4.26	18.1476	2.06398	6.52687
4.27	18.2329	2.06640	6.53452
4.28	18.3184	2.06882	6.54217
4.29	18.4041	2.07123	6.54981
4.30	18.4900	2.07364	6.55744
4.31	18.5761	2.07605	6.56506
4.32	18.6624	2.07846	6.57267
4.33	18.7489	2.08087	6.58027
4.34	18.8356	2.08327	6.58787
4.35	18.9225	2.08567	6.59545
4.36	19.0096	2.08806	6.60303
4.37	19.0969	2.09045	6.61060
4.38	19.1844	2.09284	6.61816
4.39	19.2721	2.09523	6.62571
4.40	19.3600	2.09762	6.63325
4.41	19.4481	2.10000	6.64078
4.42	19.5364	2.10238	6.64831
4.43	19.6249	2.10476	6.65582
4.44	19.7136	2.10713	6.66333
4.45	19.8025	2.10950	6.67083
4.46	19.8916	2.11187	6.67832
4.47	19.9809	2.11424	6.68581
4.48	20.0704	2.11660	6.69328
4.49	20.1601	2.11896	6.70075
4.50	20.2500	2.12132	6.70820
N	N ²	√N	√10N

N	N ²	√N	√10N
4.50	20.2500	2.12132	6.70820
4.51	20.3401	2.12368	6.71565
4.52	20.4304	2.12603	6.72309
4.53	20.5209	2.12838	6.73053
4.54	20.6116	2.13073	6.73795
4.55	20.7025	2.13307	6.74537
4.56	20.7936	2.13542	6.75278
4.57	20.8849	2.13776	6.76018
4.58	20.9764	2.14009	6.76757
4.59	21.0681	2.14243	6.77495
4.60	21.1600	2.14476	6.78233
4.61	21.2521	2.14709	6.78970
4.62	21.3444	2.14942	6.79706
4.63	21.4369	2.15174	6.80441
4.64	21.5296	2.15407	6.81175
4.65	21.6225	2.15639	6.81909
4.66	21.7156	2.15870	6.82642
4.67	21.8089	2.16102	6.83374
4.68	21.9024	2.16333	6.84105
4.69	21.9961	2.16564	6.84836
4.70	22.0900	2.16795	6.85565
4.71	22.1841	2.17025	6.86294
4.72	22.2784	2.17256	6.87023
4.73	22.3729	2.17486	6.87750
4.74	22.4676	2.17715	6.88477
4.75	22.5625	2.17945	6.89202
4.76	22.6576	2.18174	6.89928
4.77	22.7529	2.18403	6.90652
4.78	22.8484	2.18632	6.91375
4.79	22.9441	2.18861	6.92098
4.80	23.0400	2.19089	6.92820
4.81	23.1361	2.19317	6.93542
4.82	23.2324	2.19545	6.94262
4.83	23.3289	2.19773	6.94982
4.84	23.4256	2.20000	6.95701
4.85	23.5225	2.20227	6.96419
4.86	23.6196	2.20454	6.97137
4.87	23.7169	2.20681	6.97854
4.88	23.8144	2.20907	6.98570
4.89	23.9121	2.21133	6.99285
4.90	24.0100	2.21359	7.00000
4.91	24.1081	2.21585	7.00714
4.92	24.2064	2.21811	7.01427
4.93	24.3049	2.22036	7.02140
4.94	24.4036	2.22261	7.02851
4.95	24.5025	2.22486	7.03562
4.96	24.6016	2.22711	7.04273
4.97	24.7009	2.22935	7.04982
4.98	24.8004	2.23159	7.05691
4.99	24.9001	2.23383	7.06399
5.00	25.0000	2.23607	7.07107
N	N ²	√N	√10N

SQUARES AND SQUARE ROOTS—(Continued)

N	N²	√N	√10N
5.00	25.0000	2.23607	7.07107
5.01	25.1001	2.23830	7.07814
5.02	25.2004	2.24054	7.08520
5.03	25.3009	2.24277	7.09225
5.04	25.4016	2.24499	7.09930
5.05	25.5025	2.24722	7.10634
5.06	25.6036	2.24944	7.11337
5.07	25.7049	2.25167	7.12039
5.08	25.8064	2.25389	7.12741
5.09	25.9081	2.25610	7.13442
5.10	26.0100	2.25832	7.14143
5.11	26.1121	2.26053	7.14843
5.12	26.2144	2.26274	7.15542
5.13	26.3169	2.26495	7.16240
5.14	26.4196	2.26716	7.16938
5.15	26.5225	2.26936	7.17635
5.16	26.6256	2.27156	7.18331
5.17	26.7289	2.27376	7.19027
5.18	26.8324	2.27596	7.19722
5.19	26.9361	2.27816	7.20417
5.20	27.0400	2.28035	7.21110
5.21	27.1441	2.28254	7.21803
5.22	27.2484	2.28473	7.22496
5.23	27.3529	2.28692	7.23187
5.24	27.4576	2.28910	7.23878
5.25	27.5625	2.29129	7.24569
5.26	27.6676	2.29347	7.25259
5.27	27.7729	2.29565	7.25948
5.28	27.8784	2.29783	7.26636
5.29	27.9841	2.30000	7.27324
5.30	28.0900	2.30217	7.28011
5.31	28.1961	2.30434	7.28697
5.32	28.3024	2.30651	7.29383
5.33	28.4089	2.30868	7.30068
5.34	28.5156	2.31084	7.30753
5.35	28.6225	2.31301	7.31437
5.36	28.7296	2.31517	7.32120
5.37	28.8369	2.31733	7.32803
5.38	28.9444	2.31948	7.33485
5.39	29.0521	2.32164	7.34166
5.40	29.1600	2.32379	7.34847
5.41	29.2681	2.32594	7.35527
5.42	29.3764	2.32809	7.36206
5.43	29.4849	2.33024	7.36885
5.44	29.5936	2.33238	7.37564
5.45	29.7025	2.33452	7.38241
5.46	29.8116	2.33666	7.38918
5.47	29.9209	2.33880	7.39594
5.48	30.0304	2.34094	7.40270
5.49	30.1401	2.34307	7.40945
5.50	30.2500	2.34521	7.41620
N	N²	√N	√10N

N	N²	√N	√10N
5.50	30.2500	2.34521	7.41620
5.51	30.3601	2.34734	7.42294
5.52	30.4704	2.34947	7.42967
5.53	30.5809	2.35160	7.43640
5.54	30.6916	2.35372	7.44312
5.55	30.8025	2.35584	7.44983
5.56	30.9136	2.35797	7.45654
5.57	31.0249	2.36008	7.46324
5.58	31.1364	2.36220	7.46994
5.59	31.2481	2.36432	7.47663
5.60	31.3600	2.36643	7.48331
5.61	31.4721	2.36854	7.48999
5.62	31.5844	2.37065	7.49667
5.63	31.6969	2.37276	7.50333
5.64	31.8096	2.37487	7.50999
5.65	31.9225	2.37697	7.51665
5.66	32.0356	2.37908	7.52330
5.67	32.1489	2.38118	7.52994
5.68	32.2624	2.38328	7.53658
5.69	32.3761	2.38537	7.54321
5.70	32.4900	2.38747	7.54983
5.71	32.6041	2.38956	7.55645
5.72	32.7184	2.39165	7.56307
5.73	32.8329	2.39374	7.56968
5.74	32.9476	2.39583	7.57628
5.75	33.0625	2.39792	7.58288
5.76	33.1776	2.40000	7.58947
5.77	33.2929	2.40208	7.59605
5.78	33.4084	2.40416	7.60263
5.79	33.5241	2.40624	7.60920
5.80	33.6400	2.40832	7.61577
5.81	33.7561	2.41039	7.62234
5.82	33.8724	2.41247	7.62889
5.83	33.9889	2.41454	7.63544
5.84	34.1056	2.41661	7.64199
5.85	34.2225	2.41868	7.64853
5.86	34.3396	2.42074	7.65506
5.87	34.4569	2.42281	7.66159
5.88	34.5744	2.42487	7.66812
5.89	34.6921	2.42693	7.67463
5.90	34.8100	2.42899	7.68115
5.91	34.9281	2.43105	7.68765
5.92	35.0464	2.43311	7.69415
5.93	35.1649	2.43516	7.70065
5.94	35.2836	2.43721	7.70714
5.95	35.4025	2.43926	7.71362
5.96	35.5216	2.44131	7.72010
5.97	35.6409	2.44336	7.72658
5.98	35.7604	2.44540	7.73305
5.99	35.8801	2.44745	7.73951
6.00	36.0000	2.44949	7.74597
N	N²	√N	√10N

SQUARES AND SQUARE ROOTS—(Continued)

N	N²	√N	√10N
6.00	36.0000	2.44949	7.74597
6.01	36.1201	2.45153	7.75242
6.02	36.2404	2.45357	7.75887
6.03	36.3609	2.45561	7.76531
6.04	36.4816	2.45764	7.77174
6.05	36.6025	2.45967	7.77817
6.06	36.7236	2.46171	7.78460
6.07	36.8449	2.46374	7.79102
6.08	36.9664	2.46577	7.79744
6.09	37.0881	2.46779	7.80385
6.10	37.2100	2.46982	7.81025
6.11	37.3321	2.47184	7.81665
6.12	37.4544	2.47386	7.82304
6.13	37.5769	2.47588	7.82943
6.14	37.6996	2.47790	7.83582
6.15	37.8225	2.47992	7.84219
6.16	37.9456	2.48193	7.84857
6.17	38.0689	2.48395	7.85493
6.18	38.1924	2.48596	7.86130
6.19	38.3161	2.48797	7.86766
6.20	38.4400	2.48998	7.87401
6.21	38.5641	2.49199	7.88036
6.22	38.6884	2.49399	7.88670
6.23	38.8129	2.49600	7.89303
6.24	38.9376	2.49800	7.89937
6.25	39.0625	2.50000	7.90569
6.26	39.1876	2.50200	7.91202
6.27	39.3129	2.50400	7.91833
6.28	39.4384	2.50599	7.92465
6.29	39.5641	2.50799	7.93095
6.30	39.6900	2.50998	7.93725
6.31	39.8161	2.51197	7.94355
6.32	39.9424	2.51396	7.94984
6.33	40.0689	2.51595	7.95613
6.34	40.1956	2.51794	7.96241
6.35	40.3225	2.51992	7.96869
6.36	40.4496	2.52190	7.97496
6.37	40.5769	2.52389	7.98123
6.38	40.7044	2.52587	7.98749
6.39	40.8321	2.52784	7.99375
6.40	40.9600	2.52982	8.00000
6.41	41.0881	2.53180	8.00625
6.42	41.2164	2.53377	8.01249
6.43	41.3449	2.53574	8.01873
6.44	41.4736	2.53772	8.02496
6.45	41.6025	2.53969	8.03119
6.46	41.7316	2.54165	8.03741
6.47	41.8609	2.54362	8.04363
6.48	41.9904	2.54558	8.04984
6.49	42.1201	2.54755	8.05605
6.50	42.2500	2.54951	8.06226
N	N²	√N	√10N

N	N²	√N	√10N
6.50	42.2500	2.54951	8.06226
6.51	42.3801	2.55147	8.06846
6.52	42.5104	2.55343	8.07465
6.53	42.6409	2.55539	8.08084
6.54	42.7716	2.55734	8.08703
6.55	42.9025	2.55930	8.09321
6.56	43.0336	2.56125	8.09938
6.57	43.1649	2.56320	8.10555
6.58	43.2964	2.56515	8.11172
6.59	43.4281	2.56710	8.11788
6.60	43.5600	2.56905	8.12404
6.61	43.6921	2.57099	8.13019
6.62	43.8244	2.57294	8.13634
6.63	43.9569	2.57488	8.14248
6.64	44.0896	2.57682	8.14862
6.65	44.2225	2.57876	8.15475
6.66	44.3556	2.58070	8.16088
6.67	44.4889	2.58263	8.16701
6.68	44.6224	2.58457	8.17313
6.69	44.7561	2.58650	8.17924
6.70	44.8900	2.58844	8.18535
6.71	45.0241	2.59037	8.19146
6.72	45.1584	2.59230	8.19756
6.73	45.2929	2.59422	8.20366
6.74	45.4276	2.59615	8.20975
6.75	45.5625	2.59808	8.21584
6.76	45.6976	2.60000	8.22192
6.77	45.8329	2.60192	8.22800
6.78	45.9684	2.60384	8.23408
6.79	46.1041	2.60576	8.24015
6.80	46.2400	2.60768	8.24621
6.81	46.3761	2.60960	8.25227
6.82	46.5124	2.61151	8.25833
6.83	46.6489	2.61343	8.26438
6.84	46.7856	2.61534	8.27043
6.85	46.9225	2.61725	8.27647
6.86	47.0596	2.61916	8.28251
6.87	47.1969	2.62107	8.28855
6.88	47.3344	2.62298	8.29458
6.89	47.4721	2.62488	8.30060
6.90	47.6100	2.62679	8.30662
6.91	47.7481	2.62869	8.31264
6.92	47.8864	2.63059	8.31865
6.93	48.0249	2.63249	8.32466
6.94	48.1636	2.63439	8.33067
6.95	48.3025	2.63629	8.33667
6.96	48.4416	2.63818	8.34266
6.97	48.5809	2.64008	8.34865
6.98	48.7204	2.64197	8.35464
6.99	48.8601	2.64386	8.36062
7.00	49.0000	2.64575	8.36660
N	N²	√N	√10N

SQUARES AND SQUARE ROOTS—(Continued)

N	N ²	√N	√10N
7.00	49.0000	2.64575	8.36660
7.01	49.1401	2.64764	8.37257
7.02	49.2804	2.64953	8.37854
7.03	49.4209	2.65141	8.38451
7.04	49.5616	2.65330	8.39047
7.05	49.7025	2.65518	8.39643
7.06	49.8436	2.65707	8.40238
7.07	49.9849	2.65895	8.40833
7.08	50.1264	2.66083	8.41427
7.09	50.2681	2.66271	8.42021
7.10	50.4100	2.66458	8.42615
7.11	50.5521	2.66646	8.43208
7.12	50.6944	2.66833	8.43801
7.13	50.8369	2.67021	8.44393
7.14	50.9796	2.67208	8.44985
7.15	51.1225	2.67395	8.45577
7.16	51.2656	2.67582	8.46168
7.17	51.4089	2.67769	8.46759
7.18	51.5524	2.67955	8.47349
7.19	51.6961	2.68142	8.47939
7.20	51.8400	2.68328	8.48528
7.21	51.9841	2.68514	8.49117
7.22	52.1284	2.68701	8.49706
7.23	52.2729	2.68887	8.50294
7.24	52.4176	2.69072	8.50882
7.25	52.5625	2.69258	8.51469
7.26	52.7076	2.69444	8.52056
7.27	52.8529	2.69629	8.52643
7.28	52.9984	2.69815	8.53229
7.29	53.1441	2.70000	8.53815
7.30	53.2900	2.70185	8.54400
7.31	53.4361	2.70370	8.54985
7.32	53.5824	2.70555	8.55570
7.33	53.7289	2.70740	8.56154
7.34	53.8756	2.70924	8.56738
7.35	54.0225	2.71109	8.57321
7.36	54.1696	2.71293	8.57904
7.37	54.3169	2.71477	8.58487
7.38	54.4644	2.71662	8.59069
7.39	54.6121	2.71846	8.59651
7.40	54.7600	2.72029	8.60233
7.41	54.9081	2.72213	8.60814
7.42	55.0564	2.72397	8.61394
7.43	55.2049	2.72580	8.61974
7.44	55.3536	2.72764	8.62554
7.45	55.5025	2.72947	8.63134
7.46	55.6516	2.73130	8.63713
7.47	55.8009	2.73313	8.64292
7.48	55.9504	2.73496	8.64870
7.49	56.1001	2.73679	8.65448
7.50	56.2500	2.73861	8.66025
N	N²	√N	√10N

N	N ²	√N	√10N
7.50	56.2500	2.73861	8.66025
7.51	56.4001	2.74044	8.66603
7.52	56.5504	2.74226	8.67179
7.53	56.7009	2.74408	8.67756
7.54	56.8516	2.74591	8.68332
7.55	57.0025	2.74773	8.68907
7.56	57.1536	2.74955	8.69483
7.57	57.3049	2.75136	8.70057
7.58	57.4564	2.75318	8.70632
7.59	57.6081	2.75500	8.71206
7.60	57.7600	2.75681	8.71780
7.61	57.9121	2.75862	8.72353
7.62	58.0644	2.76043	8.72926
7.63	58.2169	2.76225	8.73499
7.64	58.3696	2.76405	8.74071
7.65	58.5225	2.76586	8.74643
7.66	58.6756	2.76767	8.75214
7.67	58.8289	2.76948	8.75785
7.68	58.9824	2.77128	8.76356
7.69	59.1361	2.77308	8.76926
7.70	59.2900	2.77489	8.77496
7.71	59.4441	2.77669	8.78066
7.72	59.5984	2.77849	8.78635
7.73	59.7529	2.78029	8.79204
7.74	59.9076	2.78209	8.79773
7.75	60.0625	2.78388	8.80341
7.76	60.2176	2.78568	8.80909
7.77	60.3729	2.78747	8.81476
7.78	60.5284	2.78927	8.82043
7.79	60.6841	2.79106	8.82610
7.80	60.8400	2.79285	8.83176
7.81	60.9961	2.79464	8.83742
7.82	61.1524	2.79643	8.84308
7.83	61.3089	2.79821	8.84873
7.84	61.4656	2.80000	8.85438
7.85	61.6225	2.80179	8.86002
7.86	61.7796	2.80357	8.86566
7.87	61.9369	2.80535	8.87130
7.88	62.0944	2.80713	8.87694
7.89	62.2521	2.80891	8.88257
7.90	62.4100	2.81069	8.88819
7.91	62.5681	2.81247	8.89382
7.92	62.7264	2.81425	8.89944
7.93	62.8849	2.81603	8.90505
7.94	63.0436	2.81780	8.91067
7.95	63.2025	2.81957	8.91628
7.96	63.3616	2.82135	8.92188
7.97	63.5209	2.82312	8.92749
7.98	63.6804	2.82489	8.93308
7.99	63.8401	2.82666	8.93868
8.00	64.0000	2.82843	8.94427
N	N²	√N	√10N

SQUARES AND SQUARE ROOTS—(Continued)

N	N²	√N	√10N
8.00	64.0000	2.82843	8.94427
8.01	64.1601	2.83019	8.94986
8.02	64.3204	2.83196	8.95545
8.03	64.4809	2.83373	8.96103
8.04	64.6416	2.83549	8.96660
8.05	64.8025	2.83725	8.97218
8.06	64.9636	2.83901	8.97775
8.07	65.1249	2.84077	8.98332
8.08	65.2864	2.84253	8.98888
8.09	65.4481	2.84429	8.99444
8.10	65.6100	2.84605	9.00000
8.11	65.7721	2.84781	9.00555
8.12	65.9344	2.84956	9.01110
8.13	66.0969	2.85132	9.01665
8.14	66.2596	2.85307	9.02219
8.15	66.4225	2.85482	9.02774
8.16	66.5856	2.85657	9.03327
8.17	66.7489	2.85832	9.03881
8.18	66.9124	2.86007	9.04434
8.19	67.0761	2.86182	9.04986
8.20	67.2400	2.86356	9.05539
8.21	67.4041	2.86531	9.06091
8.22	67.5684	2.86705	9.06642
8.23	67.7329	2.86880	9.07193
8.24	67.8976	2.87054	9.07744
8.25	68.0625	2.87228	9.08295
8.26	68.2276	2.87402	9.08845
8.27	68.3929	2.87576	9.09395
8.28	68.5584	2.87750	9.09945
8.29	68.7241	2.87924	9.10494
8.30	68.8900	2.88097	9.11043
8.31	69.0561	2.88271	9.11592
8.32	69.2224	2.88444	9.12140
8.33	69.3889	2.88617	9.12688
8.34	69.5556	2.88791	9.13236
8.35	69.7225	2.88964	9.13783
8.36	69.8896	2.89137	9.14330
8.37	70.0569	2.89310	9.14877
8.38	70.2244	2.89482	9.15423
8.39	70.3921	2.89655	9.15969
8.40	70.5600	2.89828	9.16515
8.41	70.7281	2.90000	9.17061
8.42	70.8964	2.90172	9.17606
8.43	71.0649	2.90345	9.18150
8.44	71.2336	2.90517	9.18695
8.45	71.4025	2.90689	9.19239
8.46	71.5716	2.90861	9.19783
8.47	71.7409	2.91033	9.20326
8.48	71.9104	2.91204	9.20869
8.49	72.0801	2.91376	9.21412
8.50	72.2500	2.91548	9.21954
N	N²	√N	√10N

N	N²	√N	√10N
8.50	72.2500	2.91548	9.21954
8.51	72.4201	2.91719	9.22497
8.52	72.5904	2.91890	9.23038
8.53	72.7609	2.92062	9.23580
8.54	72.9316	2.92233	9.24121
8.55	73.1025	2.92404	9.24662
8.56	73.2736	2.92575	9.25203
8.57	73.4449	2.92746	9.25743
8.58	73.6164	2.92916	9.26283
8.59	73.7881	2.93087	9.26823
8.60	73.9600	2.93258	9.27362
8.61	74.1321	2.93428	9.27901
8.62	74.3044	2.93598	9.28440
8.63	74.4769	2.93769	9.28978
8.64	74.6496	2.93939	9.29516
8.65	74.8225	2.94109	9.30054
8.66	74.9956	2.94279	9.30591
8.67	75.1689	2.94449	9.31128
8.68	75.3424	2.94618	9.31665
8.69	75.5161	2.94788	9.32202
8.70	75.6900	2.94958	9.32738
8.71	75.8641	2.95127	9.33274
8.72	76.0384	2.95296	9.33809
8.73	76.2129	2.95466	9.34345
8.74	76.3876	2.95635	9.34880
8.75	76.5625	2.95804	9.35414
8.76	76.7376	2.95973	9.35949
8.77	76.9129	2.96142	9.36483
8.78	77.0884	2.96311	9.37017
8.79	77.2641	2.96479	9.37550
8.80	77.4400	2.96648	9.38083
8.81	77.6161	2.96816	9.38616
8.82	77.7924	2.96985	9.39149
8.83	77.9689	2.97153	9.39681
8.84	78.1456	2.97321	9.40213
8.85	78.3225	2.97489	9.40744
8.86	78.4996	2.97658	9.41276
8.87	78.6769	2.97825	9.41807
8.88	78.8544	2.97993	9.42338
8.89	79.0321	2.98161	9.42868
8.90	79.2100	2.98329	9.43398
8.91	79.3881	2.98496	9.43928
8.92	79.5664	2.98664	9.44458
8.93	79.7449	2.98831	9.44987
8.94	79.9236	2.98998	9.45516
8.95	80.1025	2.99166	9.46044
8.96	80.2816	2.99333	9.46573
8.97	80.4609	2.99500	9.47101
8.98	80.6404	2.99666	9.47629
8.99	80.8201	2.99833	9.48156
9.00	81.0000	3.00000	9.48683
N	N²	√N	√10N

SQUARES AND SQUARE ROOTS—(Continued)

N	N²	√N	√10N
9.00	81.0000	3.00000	9.48683
9.01	81.1801	3.00167	9.49210
9.02	81.3604	3.00333	9.49737
9.03	81.5409	3.00500	9.50263
9.04	81.7216	3.00666	9.50789
9.05	81.9025	3.00832	9.51315
9.06	82.0836	3.00998	9.51840
9.07	82.2649	3.01164	9.52365
9.08	82.4464	3.01330	9.52890
9.09	82.6281	3.01496	9.53415
9.10	82.8100	3.01662	9.53939
9.11	82.9921	3.01828	9.54463
9.12	83.1744	3.01993	9.54987
9.13	83.3569	3.02159	9.55510
9.14	83.5396	3.02324	9.56033
9.15	83.7225	3.02490	9.56556
9.16	83.9056	3.02655	9.57079
9.17	84.0889	3.02820	9.57601
9.18	84.2724	3.02985	9.58123
9.19	84.4561	3.03150	9.58645
9.20	84.6400	3.03315	9.59166
9.21	84.8241	3.03480	9.59687
9.22	85.0084	3.03645	9.60208
9.23	85.1929	3.03809	9.60729
9.24	85.3776	3.03974	9.61249
9.25	85.5625	3.04138	9.61769
9.26	85.7476	3.04302	9.62289
9.27	85.9329	3.04467	9.62808
9.28	86.1184	3.04631	9.63328
9.29	86.3041	3.04795	9.63846
9.30	86.4900	3.04959	9.64365
9.31	86.6761	3.05123	9.64883
9.32	86.8624	3.05287	9.65401
9.33	87.0489	3.05450	9.65919
9.34	87.2356	3.05614	9.66437
9.35	87.4225	3.05778	9.66954
9.36	87.6096	3.05941	9.67471
9.37	87.7969	3.06105	9.67988
9.38	87.9844	3.06268	9.68504
9.39	88.1721	3.06431	9.69020
9.40	88.3600	3.06594	9.69536
9.41	88.5481	3.06757	9.70052
9.42	88.7364	3.06920	9.70567
9.43	88.9249	3.07083	9.71082
9.44	89.1136	3.07246	9.71597
9.45	89.3025	3.07409	9.72111
9.46	89.4916	3.07571	9.72625
9.47	89.6809	3.07734	9.73139
9.48	89.8704	3.07896	9.73653
9.49	90.0601	3.08058	9.74166
9.50	90.2500	3.08221	9.74679
N	N²	√N	√10N

N	N²	√N	√10N
9.50	90.2500	3.08221	9.74679
9.51	90.4401	3.08383	9.75192
9.52	90.6304	3.08545	9.75705
9.53	90.8209	3.08707	9.76217
9.54	91.0116	3.08869	9.76729
9.55	91.2025	3.09031	9.77241
9.56	91.3936	3.09192	9.77753
9.57	91.5849	3.09354	9.78264
9.58	91.7764	3.09516	9.78775
9.59	91.9681	3.09677	9.79285
9.60	92.1600	3.09839	9.79796
9.61	92.3521	3.10000	9.80306
9.62	92.5444	3.10161	9.80816
9.63	92.7369	3.10322	9.81326
9.64	92.9296	3.10483	9.81835
9.65	93.1225	3.10644	9.82344
9.66	93.3156	3.10805	9.82853
9.67	93.5089	3.10966	9.83362
9.68	93.7024	3.11127	9.83870
9.69	93.8961	3.11288	9.84378
9.70	94.0900	3.11448	9.84886
9.71	94.2841	3.11609	9.85393
9.72	94.4784	3.11769	9.85901
9.73	94.6729	3.11929	9.86408
9.74	94.8676	3.12090	9.86914
9.75	95.0625	3.12250	9.87421
9.76	95.2576	3.12410	9.87927
9.77	95.4529	3.12570	9.88433
9.78	95.6484	3.12730	9.88939
9.79	95.8441	3.12890	9.89444
9.80	96.0400	3.13050	9.89949
9.81	96.2361	3.13209	9.90454
9.82	96.4324	3.13369	9.90959
9.83	96.6289	3.13528	9.91464
9.84	96.8256	3.13688	9.91968
9.85	97.0225	3.13847	9.92472
9.86	97.2196	3.14006	9.92975
9.87	97.4169	3.14166	9.93479
9.88	97.6144	3.14325	9.93982
9.89	97.8121	3.14484	9.94485
9.90	98.0100	3.14643	9.94987
9.91	98.2081	3.14802	9.95490
9.92	98.4064	3.14960	9.95992
9.93	98.6049	3.15119	9.96494
9.94	98.8036	3.15278	9.96995
9.95	99.0025	3.15436	9.97497
9.96	99.2016	3.15595	9.97998
9.97	99.4009	3.15753	9.98499
9.98	99.6004	3.15911	9.98999
9.99	99.8001	3.16070	9.99500
10.00	100.000	3.16228	10.0000
N	N²	√N	√10N

APPENDIX

Answers to Problems

In the event that your answers to the computed values of the statistical tests approximate, but do not precisely equal, those given below, you should first consider the “number of places” used in the different computations in order to understand discrepancies.

CHAPTER 5

1. With 26 *df*, a *t* of 2.14 is significant beyond the .05 level. Hence the null hypothesis may be rejected.
2. With 30 *df*, a *t* of 2.20 is significant beyond the .05 level. Since this was the criterion set for rejecting the null hypothesis, and since the direction of the means is that specified by the empirical hypothesis, it may be concluded that the empirical hypothesis was confirmed — that the independent variable influenced the dependent variable.
3. With 13 *df*, the computed *t* of 4.30 is significant beyond the 1 per cent

level. Since the group that received the tranquilizer had the lesser mean psychotic tendency it may be concluded that the drug produces the advertised effect.

4. The computed t of .51 is not significant. Since the experimenter could not reject his null hypothesis, he failed to confirm his empirical hypothesis.

5. His suspicion is not confirmed — the computed t is .10.

CHAPTER 6

1. The confounding in this study is especially atrocious. The subjects in the two groups undoubtedly differ in a large number of respects other than type of method. For instance, there may be differences in intelligence, opportunity to study, socio-economic level, as well as differences in reading proficiency prior to learning by either method, and certainly there were different teachers. The proper approach would be to randomly assign subjects from the same class in a given school to two groups, and then to randomly determine which group is taught by each method, both groups being taught by the same instructor.

2. The characteristics of the individual tanks and targets are confounded with the independent variable. It may be that one tank gun is more accurate than the other, and that one set of targets is easier to hit than the other. To control these variables one might have all subjects fire from the same tank (continually checking the calibration of the gun) on the same set of targets. Or half of the subjects from each group could fire from each tank onto each set of targets.

3. The conclusion reached in this study is limited to the effects of class from which the children came. Undoubtedly these classes differ in a number of respects, among which is age at which they are toilet trained. The dependent variable results may thus be due to some other differential experience of the groups such as amount of social stimulation, or amount of money spent on family needs. The obvious, but difficult, way to conduct this experiment in order to establish a causal relation would be to randomly select a group of children, randomly assign them to two groups, and then randomly determine the age at which each group is toilet trained.

4. The control group should also be operated on, except that the hypothalamus should not be impaired. It could be that some structure other than the hypothalamus is disturbed during the operation, and this other structure may be responsible for the "missing" behavior.

5. There may be other reasons for not reporting an emotionally loaded word than that it is not perceived. For instance "sex" may actually be perceived by a subject, but he waits until he is absolutely sure that that is the

word, possibly saving himself from a "social blunder." In addition, the frequency with which the loaded and neutral words are used in everyday life undoubtedly differs, thus affecting the threshold for recognition of the words. A better approach would be to start with a number of words that are emotionally neutral (or with nonsense syllables), and make some of them emotionally loaded (such as associating an electric shock with them). The loaded and neutral words should be equated for frequency of use.

CHAPTER 8

1. With 7 *df* the computed *t* of 2.58 is significant at the 5 per cent level. Hence the null hypothesis may be rejected. However the subjects who used the Eastern Grip had a higher mean score, from which we can conclude that the empirical hypothesis is not confirmed.

2. The computed *t* of 2.98 with 6 *df* is significant beyond the .05 level. Since the experimental group had the higher mean score, the empirical hypothesis is confirmed.

3. With 19 *df*, the computed *t* of 7.02 is significant beyond the .02 level. Since the group that used the training aid had the higher mean score, we may conclude that the training aid facilitated map reading proficiency.

CHAPTER 9

1. $R_2 = 1.04$ and $R_3 = 1.09$

Mean Scores for Groups

<i>English Majors</i>	<i>Art Majors</i>	<i>Chemistry Majors</i>
2.17	5.50	9.33

All groups are significantly different from each other.

2. $R_2 = 2.42$, $R_3 = 2.54$, and $R_4 = 2.62$.

Mean Scores for Groups

III	II	I	IV
2.00	2.38	6.25	7.12

Groups II and III both have significantly lower means than do Groups I and IV. It might be added that, for greatest proficiency, these fictitious data indicate that considerable practice or extremely little practice are most beneficial (a U-shaped curve).

3. $R_2 = 2.28$, $R_3 = 2.40$, $R_4 = 2.47$ and $R_5 = 2.53$

Mean Scores for Groups

1	2	3	4	5
3.45	3.82	4.18	6.64	8.45

The order of means increases systematically with the independent variable. The higher two means are significantly superior to the lower three means. In general, the hypothesis was confirmed. It would have been more desirable, however, to have obtained a significant difference between each of the groups, a goal that might have been achieved had a larger number of subjects per group been studied.

4. R_2 for the comparison between Groups B and C is 7.36, between Groups A and B is 8.33, and R_3 for the comparison between Groups A and C is 8.35.

Mean Scores for Groups Taught by:

Method C	Method B	Method A
27.23	29.90	46.62

Method A is to be preferred since it led to significantly greater proficiency than did the other two methods.

CHAPTER 10

1.

		Amount of drug administered	
		None	2cc.
Type of psychosis	Manic depressive		
	Schizophrenic		

2.

Analysis of Variance

<i>Source of Variation</i>	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>
Over-all Among	(91.13)	(3)		
Between Drugs	89.28	1	89.28	39.33
Between Psychoses	1.28	1	1.28	.56
D × P	.57	1	.57	.25
Within Groups	54.58	24	2.27	
Total	145.71	27		

Since the F for "between drugs" is significant, variation of this independent variable is effective. The mean score for the subjects who received drugs is higher than that for those who did not receive drugs. Hence we may conclude that administration of the drug led to an increase in normality. The lack of significant F 's for the "between psychoses" and interaction sources of variation indicates that there is no difference in normality as a function of type of psychosis, nor that there is an interaction between the variables.

3.

Amount of drug administered

		None	2cc.	4cc.
Type of psychosis	Paranoid			
	Manic depressive			
	Schizophrenic			

4.

Amount of drug administered

		None	2cc.	4cc.	6cc.
Type of subject	Paranoid				
	Manic depressive				
	Schizophrenic				
	Normal				

5.

Type of brand

		Old Zincs	Counts
Filter	Without filter		
	With filter		

6.

Analysis of Variance

<i>Source of Variation</i>	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>
Over-all Among	(275.88)	(3)		
Between Brands	0.03	1	0.03	0.02
Between Filters	0.23	1	0.23	0.12
B \times F	275.62	1	275.62	144.30
Within Groups	68.90	36	1.91	
Total	344.78	39		

Since neither variation of brands nor filters resulted in significant differences, we may conclude that variation of these variables, considered by themselves, did not affect steadiness. However the interaction was highly significant. From this we may conclude that whether or not brand affects steadiness depends on whether or not a filter was used — that smoking Old Zincs with a filter leads to greater steadiness than does smoking Counts with a filter, but that smoking Counts without a filter leads to greater steadiness than smoking Old Zincs without a filter. It would appear that putting a filter on Counts decreases steadiness, but putting a filter on Old Zincs increases steadiness. In fact, Counts without a filter leads to about the same amount of steadiness as Old Zincs with a filter, as a diagram of the interaction would show. But we don't recommend that you smoke either brand.

7.

Analysis of Variance

<i>Source of Variation</i>	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>
Over-all Among	(80.68)	(3)	26.89	
Between Opium	30.04	1	30.04	27.81
Between Marijuana	48.89	1	48.89	45.27
O \times M	1.75	1	1.75	1.62
Within Groups	26.00	24	1.08	
Total	106.68	27		

Since the *F*'s for "between Opium" and "between Marijuana" are both significant, we can conclude that smoking both of them lead to hallucinatory activity, and that there is no interaction between these two variables since this latter source of variation is not significant.

REFERENCES

Albrecht, Heather A. Replication of pronoun satiation by prior verbal stimulation. Unpublished manuscript, 1965.

Albrecht, Heather A., and Webster, R. L. The effect of prior verbal stimulation on associates to stimulus words. Unpublished manuscript, 1966.

Anderson, R. L., and Bancroft, T. A. *Statistical theory in research*. New York: McGraw-Hill, 1952.

Asher, R. Why are medical journals so dull? *British Medical Journal*, 1958, II, 502.

Babich, F. R., Jacobson, A. L., Bubash, Suzanne, and Jacobson, Ann. Transfer of a response to naive rats by injection of ribonucleic acid extracted from trained rats. *Science*, 1965, 149, 656-657.

Bachrach, A. J. *Psychological research: An Introduction* (2nd. ed.). New York: Random House, 1965.

Barber, B. Resistance by scientists to scientific discovery. *Science*, 1961, 134, 596-602.

- Batkin, S.; Woodward, W. T., Cole, R. E., & Hall, J. B., RNA and actinomycin-D enhancement of learning in the carp. *Psychonomic Science*, 1966, 5(9), 345-346.
- Bersh, P. J. The influence of two variables upon the establishment of a secondary reinforcer for operant responses. *Journal of Experimental Psychology*, 1951, 41, 62-73.
- Binder A. Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, 1963, 70, 107-115.
- Binder, A., McConnell, D., and Sjolholm, N. A. Verbal conditioning as a function of experimenter characteristics. *Journal of Abnormal and Social Psychology*, 1957, 55, 309-314.
- Boe, E. E. Effect of punishment during and intensity on the extinction of an instrumental response. *Journal of Experimental Psychology*, 1966, 72, 125-131.
- Boneau, C. A. The effects of violation of assumptions underlying the *t*-test. *Psychological Bulletin*, 1960, 57, 49-64.
- Bridgeman, P. W. *The logic of modern physics*. New York: Macmillan, 1927.
- Brown, C. C., and Saucer, R. T. *Electronic instrumentation for the behavioral sciences*. Springfield, Illinois: Charles C Thomas, 1958.
- Brunswik, E. *Perception and the representative design of psychological experiments*. Berkeley and Los Angeles: University of California Press, 1956.
- Bugelski, B. R. *Experimental psychology*. New York: Henry Holt, 1951.
- Byrne, W. L., et al. Memory transfer. *Science*, 1966, 153, 658-659.
- Burch, M. E., and Magsdick, W. K. The retention in rats of an incompletely learned maze solution for short intervals of time. *Journals of Comparative Psychology*, 1933, 16, 385-409.
- Calvin, A. D., Scriven, M., Gallagher, J. J., Hanley, C., McConnell, J. V., and McGuigan, F. J. *Psychology*. Boston: Allyn & Bacon, 1961.
- Cannon, W. B. *The way of an investigator*. New York: Norton, 1945.
- Cantril, H. *The invasion from Mars*. Princeton: Princeton University Press, 1940.
- Cochran, W. G., and Cox, G. M. *Experimental designs*. New York: Wiley, 1957.
- Cohen, M. R., and Nagel, E. *Logic and scientific method*. New York: Harcourt, Brace, 1934.
- Cook, B. S., and Hilgard, E. R. Distributed practice in motor learning: Progressively increasing and decreasing rests. *Journal of Experimental Psychology*, 1949, 39, 169-172.
- Cornsweet, T. N. *The design of electric circuits in the behavioral sciences*. New York: Wiley, 1963.

- Deese, J. *The psychology of learning*. New York: McGraw-Hill, 1952.
- Dixon, W. J., and Massey, F. J., Jr. *Statistical analysis*. New York: McGraw-Hill, 1951.
- Dore, L. R., and Hilgard, E. R. Spaced practice as a test of Snoddy's two processes in mental growth. *Journal of Experimental Psychology*, 1938, 23, 359-374.
- Doyle, A. C. *Sherlock Holmes*. Garden City, New York: Garden City Press, 1938.
- Dubs, H. H. *Rational induction*. Chicago: University of Chicago Press, 1930.
- Duncan, D. B. Multiple range and multiple F tests. *Biometrics*, 1955, 11, 1-42.
- Duncan, D. B. Multiple range tests for correlated and heteroscedastic means. *Biometrics*, 1957, 13, 164-176.
- Duncan, D. B. A simple Bayes solution to a common multiple comparisons problem. *Institute of Statistics Mimeograph Series No. 223*, University of North Carolina, April 1959.
- Ebbinghaus, E. *Memory: A contribution to experimental psychology* (Translated by H. A. Ruger and C. E. Busserius). New York: Columbia University Press, 1913.
- Edgington, E. S. A tabulation of inferential statistics used in psychology journals. *American Psychologist*, 1964, 19, 202-203.
- Edwards, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
- Edwards, A. L. *Experimental design in psychological research*, 3rd ed. New York: Holt, Rinehart & Winston, 1968.
- Edwards, W. Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, 1965, 63, 400-402.
- Engram, W. C. One aspect of the social psychology of experimentation: The E effect and related personality characteristics in the experimenter. Unpublished Ph.D. thesis, Cornell University, 1966.
- Farber, I. E., and Spence, K. W. Complex learning and conditioning as a function of anxiety. *Journal of Experimental Psychology*, 45, 1953, 120-125.
- Feigl, H., and Scriven, M. *Minnesota studies in the philosophy of science*. (Volume I: The foundations of science and the concepts of psychology and psychoanalysis.) Minneapolis: University of Minnesota Press, 1956.
- Fisher, A. E. Chemical stimulation of the brain. *Psychobiology*. San Francisco: W. H. Freeman & Company, 1964.
- Fisher, R. A. *The design of experiments* (6th ed.). New York: Hafner, 1953.
- Frank, P. G., ed. *The validation of scientific theories*. Boston: Beacon, 1956.

- Gaito, J. Statistical dangers involved in counterbalancing. *Psychological Reports*, 1958, 4, 463-468.
- Gaito, J. Multiple comparisons in analysis of variance. *Psychological Bulletin*, 1959, 56.
- Gaito, J. Repeated measurements designs and counterbalancing. *Psychological Bulletin*, 58, 1961, 46-54.
- Gampel, Dorothy H. Temporal factors in verbal satiation. *Journal of Experimental Psychology*, 1966, 72, 201-206.
- Grant, D. A. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, 1962, 69, 54-61.
- Grant, D. A. Personal communication, 1967.
- Grings, W. W. *Laboratory instrumentation in psychology*. Palo Alto, California: National Press, 1954.
- Grice, G. R., and Hunter, J. J. Stimulus intensity effects depend upon the type of experimental design. *Psychological Review*, 1964, 71, 247-256.
- Guilford, J. P. *Fundamental statistics in psychology and education* (4th ed.). New York: McGraw-Hill, 1965.
- Guthrie, E. R. *The psychology of learning*, Rev. ed. New York: Harper, 1952.
- Hammond, K. R. Subject and object sampling: A note. *Psychological Bulletin*, 1948, 45, 530-533.
- Hammond, K. R. Representative vs. systematic design in clinical psychology. *Psychological Bulletin*, 1954, 51, 150-159.
- Harley, W. F. Jr., and Harley, W. F. Sr. The effect of hypnosis upon paired-associate learning using an absolute method. Unpublished manuscript, 1966.
- Harlow, H. F., and Zimmerman, R. R. The development of affectional responses in infant monkeys. *American Philosophical Society*, 1958, 102, 501-509.
- Harris, F. R., Wolf, M. M., and Baer, D. M. Effects of adult social reinforcement on child behavior. *Young Children*, 1964, 20, 8-17.
- Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.
- Hempel, C. G. Studies in the logic of confirmation. *Mind*, 1945, 54, 1-26, 97-121.
- Hempel, C. G. *Aspects of scientific explanations: And other essays in the philosophy of science*. New York: Free Press, 1965.
- Hempel, C. G., and Oppenheim, P. The logic of explanation. *Philosophy of science*, 1948, 15, 135-175.

- Hernández-Péon, R., Scherrer, H., and Jouvet, M. Modification of electric activity in cochlear nucleus during "attention" in unanesthetized cats. *Science*, 1956, 123, 331-332.
- Hess, E. H. Attitude and pupil size. *Scientific American*, 1965, 212, 46-54.
- Horsnell, G. The effect of unequal group variances on the *F*-test for the homogeneity of group means. *Biometrika*, 1953, 46, 128-136.
- Hull, C. L. *Principles of behavior*. New York: Appleton-Century, 1943.
- Hull, C. L. *A behavior system*. New Haven: Yale University Press, 1952.
- Igel, G. J., and Calvin, A. D. The development of affectional responses in infant dogs. *Journal of Comparative & Physiological Psychology*, 1960, 53, 302-305.
- Jacobson, A. L., Fried, C., and Horowitz, S. D. Planarians and memory. *Nature*, 1966, 209, 599-601.
- James, W. *Principles of Psychology*. Chicago: Encyclopaedia Britannica, Inc., 1952, 809.
- Jenkins, J. G., and Dallenbach, K. M. Oblivescence during sleep and waking. *American Journal of Psychology*, 1924, 35, 605-612.
- Johnson, P., and Bailey, D. E. Some determinants of the use of relationships in discrimination learning. *Journal of Experimental Psychology*, 1966, 71, 365-372.
- Johnson, R. W. Retain the original data! *American Psychologist*, 1964, 19, 350-351.
- Jones, F. P. Experimental method in antiquity. *American Psychologist*, 1964, 19, 419-420.
- Kiesler, D. J. Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin*, 1966, 65, 110-136.
- Kramer, C. Y. Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, 1956, 12, 307-310.
- Kramer, C. Y. Extension of multiple range tests to group correlated adjusted means. *Biometrics*, 1957, 13, 13-18.
- Lachman, R., Meehan, J. T., and Bradley, Rosalee. Observing response and word association in concept shifts: Two-choice and four-choice selective learning. *Journal of Psychology*, 1965, 59, 349-357.
- Lepley, W. M. The participation of implicit speech in acts of writing. *American Journal of Psychology*, 1952, 65, 597-599.
- Li, J. C. R. *Introduction to statistical inference*. Ann Arbor, Michigan: Edwards Brothers, 1957.
- Lindquist, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton, 1953.

- McGuigan, F. J. Confirmation of theories in psychology. *Psychological Review*, 1956, 63, 98-104.
- McGuigan, F. J. The effect of precision, delay and schedule of knowledge of results on performance. *Journal of Experimental Psychology*, 1959, 58, 79-84.
- McGuigan, F. J. The experimenter: a neglected stimulus object. *Psychological Bulletin*, 1963, 60, 421-428.
- McGuigan, F. J. Covert oral behavior and auditory hallucinations. *Psychophysiology*, 1966, 3, 73-80.
- McGuigan, F. J. *Thinking: Studies of covert language processes*. New York: Appleton, 1966.
- McGuigan, F. J., Calvin, A. D., and Richardson, Elizabeth C. Manifest anxiety, palmar perspiration index, and stylus maze-learning. *American Journal of Psychology*, 1959, 67, 434-438.
- McGuigan, F. J., Crandell, Susan A., and Suiter, R. D. Covert response measures as a function of stimulus input. Paper presented at meeting of The Psychonomic Society, October 1966.
- McGuigan, F. J., Hutchens, Carolyn, Eason, Nancy and Reynolds, Teddy. The retroactive interference of motor activity with knowledge of results. *Journal of General Psychology*, 1964, 70, 279-281.
- McGuigan, F. J., and MacCaslin, E. F. Whole and part methods in learning a perceptual motor skill. *American Journal of Psychology*, 1955, 47, 658-661. (a)
- McGuigan, F. J., and MacCaslin, E. F. The relationship between rifle steadiness and rifle marksmanship and the effect of rifle training on rifle steadiness. *Journal of Applied Psychology*, 1955, 39, 156-159. (b)
- McGuigan, F. J., Ostrov, N. H., and Savukas, R. A. Covert language behavior during handwriting. In *Subvocal speech during silent reading*. Final Report, Project No. 2643, Contract No. OE J-10-073, Office of Education, 1967.
- McGuigan, F. J., and Peters, R. J., Jr. Assessing the effectiveness of programmed texts: Methodology and some findings. *Journal of Programmed Instruction*, 1965, 3, 23-34.
- McKeachie, W. J., Pollie, D., and Speisman, J. Relieving anxiety in classroom examinations. *Journal of Abnormal & Social Psychology*, 1955, 50, 93-98.
- McNemar, Q. *Psychological Statistics* (3rd ed.). New York: Wiley, 1962.
- Morgan, C. L. *An introduction to comparative psychology* (2nd ed.). London: Walter Scott, 1906.
- Mosteller, F., and Bush, R. R. Selected quantitative techniques. In G. Lindzey, ed. *Handbook of social psychology*. Cambridge, Mass.: Addison-Wesley, 1954.

- Mowrer, O. H. *Psychotherapy*. New York: Ronald, 1953.
- Newbury, E. Current interpretation and significance of Lloyd Morgan's canon. *Psychological Bulletin*, 1954, 51, 70-74.
- Overall, J. E., and Dalal, S. N. Design of experiments to maximize power relative to cost. *Psychological Bulletin*, 1965, 5, 339-350.
- Page, I. H. Serotonin. *Scientific American*, 1957, 197, 52-56.
- Peters, C. C., and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.
- Poulton, E. C., and Freeman, P. R. Unwanted asymmetrical transfer effects with balanced experimental designs. *Psychological Bulletin*, 1966, 66, 1-8.
- Ray, W. S. *An introduction to experimental design*. New York: Macmillan, 1960.
- Reichenbach, H. *Experience and prediction*. Chicago: University of Chicago Press, 1938.
- Reichenbach, H. *Elements of symbolic logic*. New York: Macmillan, 1947.
- Reichenbach, H. *The theory of probability* (2nd. ed.). Berkeley: University of California Press, 1949.
- Renshaw, S., and Schwarzbek, W. C. The dependence of the form of the pursuit meter learning function on the length of the inter-practice rests: I. Experimental. *Journal of General Psychology*, 1938, 18, 3-16.
- Reynolds, R. W., and Meeker, M. R. Thiosemicarbazide injection followed by electric shock increases resistance to stress in rats. *Science*, 1966, 151, 1101-1102.
- Rosenthal, R. The effect of the experimenter on the results of psychological research. *Progress in Experimental Personality Research*, 1964, 1, 79-114.
- Rosenthal, R. *Experimenter effects in behavioral research*. New York: Appleton, 1966.
- Ryan, T. A. Multiple comparisons in psychological research. *Psychological Bulletin*, 1959, 56, 26-47.
- Sandler, J. A test of the significance of the difference between the means of correlated measures, based on a simplification of Student's *t*. *British Journal of Psychology*, 1955, 46, 225-226.
- Schoenfeld, W. N., Antonitis, J. J., and Bersh, P. J. A preliminary study of training conditions necessary for secondary reinforcement. *Journal of Experimental Psychology*, 1950, 40, 40-45.
- Scriven, M. Explanation and prediction in evolutionary theory. *Science*, 1959, 130, 477-482.
- Seidel, R. J., and Rotberg, Iris C. Effects of written verbalization and timing of information on problem solving in programmed learning. *Journal of Educational Psychology*, 1966, 57, 151-158.

- Sidman, M. *Tactics of scientific research*. New York: Basic Books, Inc., 1960.
- Sidowski, J. B. *Experimental methods and instrumentation in psychology*. New York: McGraw-Hill, 1966.
- Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- Silverman, H. F. Effects of praise and reproof on reading growth in a non-laboratory classroom setting. *Journal of Educational Psychology*, 1957, 48, 199-206.
- Singh, S. D. Effect of human environment on cognitive behavior in the Rhesus monkey. *Journal of Comparative & Physiological Psychology*, 1966, 61, 280-283.
- Skinner, B. F. *Science and human behavior*. New York: Macmillan, 1953.
- Skinner, B. F. *Cumulative Record*. New York: Appleton, 1959.
- Solomon, R. L. An extension of control group design. *Psychological Bulletin*, 1949, 46, 137-150.
- Spence, J. T., Underwood, B. J., Duncan, C. P., and Cotton, J. W. *Elementary statistics*, Rev. ed. New York: Appleton, 1954.
- Spence, K. W. The postulates and methods of 'Behaviorism.' *Psychological Review*, 1948, 55, 67-78.
- Spence, K. W., Farber, I. E., and McFann, H. H. The relation of anxiety (drive) level to performance in competition and noncompetition paired-associates learning. *Journal of Experimental Psychology*, 1956, 52, 296-305.
- Spires, A. M. Subject-experimenter interaction in verbal conditioning. Unpublished doctoral dissertation, New York University, 1960.
- Sulzbacher, S. I. Effects of response mode and subject characteristics on learning in programmed instruction. *National Society of Programed Instruction Journal*, 1967, 4, 10-11.
- Taylor, J. A. A personality scale of manifest anxiety. *Journal of Abnormal & Social Psychology*, 1953, 48, 285-290.
- Tversky, A., and Edwards, W. Information versus reward in binary choices. *Journal of Experimental Psychology*, 1966, 71, 680-683.
- Underwood, B. J. The effect of successive interpolations on retroactive and proactive inhibition. *Psychological Monographs*, 1945, 38, 29-38.
- Underwood, B. J. *Experimental Psychology*. New York: Appleton, 1949.
- Underwood, B. J. *Psychological research*. New York: Appleton, 1957.
- Underwood, B. J. Interference and forgetting. *Psychological Review*, 1957a, 64, 49-60.
- Underwood, B. J. *Experimental psychology*. New York: Appleton, 1966.
- Underwood, B. J. *Problems in experimental design and inference*. New York: Appleton, 1966a.

Venables, P. H. and Martin, I. eds. *A manual of psychophysiological methods*. New York: Wiley, 1967.

Webster, R. L., and Weingold, H. P. Suppression of pronoun choices as a function of prior verbal stimulation. Paper read at Southeastern Psychological Association meeting, 1965.

Wilson, W., Miller, H. L., and Lower, J. S. Much ado about the null hypothesis. *Psychological Bulletin*, 1967, 69, 188-196.

Wine, R. L. *Statistics for scientists and engineers*. Englewood Cliffs: Prentice-Hall, 1964.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Wolins, L. Responsibility for raw data. *American Psychologist*, 1962, 17, 657-658.

Woods, P. J. Some characteristics of journals and authors. *American Psychologist*, 1961, 16, 699-701.

Woods, P. J., and Holland, C. H. Instrumental escape conditioning in a water tank: Effects of constant reinforcement at different points on the continuum of drive stimulus intensity. *Journal of Comparative & Physiological Psychology*, 1966, 62, 403-408.

Woodworth, R. S., and Schlosberg, H. *Experimental psychology*. New York: Henry Holt, 1955.

Yanof, H. M. *Biomedical electronics*. Philadelphia: F. A. Davis Co., 1965.

INDEX

- A*-test, statistical analysis with, 176-178
- Accuracy, of data analysis, 349-352
- Abstracts* (see *Psychological Abstracts*)
- Abstract section, experiment write-up, 91
- Adams, J. C., 33
- Adler, A., 26
- Albrecht, H. A., 18
- American Psychological Association, 64*n*, 83, 353
- American Psychologist*, 91
- "Analysis of covariance" technique, 360-361
- Analysis of Variance (table), 229
- Analysis of Variance (table), 259
- Analysis of Variance (table), 270
- Anderson, R. L., 273*n*, 355
- Antonitis, J. J., 201
- Anxiety, 173
- Apparatus, 65-70
- Apparatus description, in experiment write-up, 84
- Asher, R., 93*n*
- Assumptions, in statistical testing, 354-356
- Athenaeus, 120
- Authorship, in experiment write-up, 83
- Babich, F. R., 101
- Bachrach, A. J., 53, 78
- Bacon, F., 51
- Baer, D. M., 300-301
- Bailey, D. E., 258
- Bain, A., 110-111
- Baker, P. C., 354
- Balancing (table), 132
- Bancroft, T. A., 273*n*, 355
- Barber, B., 33
- Batkin, S., 110*n*
- Becquerel, H., 19, 37, 50
- Bell, Dr., 201*n*
- Bersh, P. J., 200-203
- Between-subjects designs, term, 289-290
- Binder, A., 315*n*, 339-340
- Boe, E. E., 255
- Boneau, C. A., 355
- Bradley, R., 266-267
- Bridgeman, P. W., 28*n*

- Brown, C. C., 67*n*
 Brunswik, E., 329*n*
 Bubash, S., 101
 Bugelski, B. R., 44
 Bunch, M. E., 29
 Bush, R. R., 353
 Byrne, W. L., 110*n*
- Calvin, A. D., 20, 110, 111, 172
 Cannon, W. B., 52
 Cantril, H., 58
 Carnap, R., 21*n*
 Case-history method, evidence report data, 57-58
 "The Case of the Mixed-up Rat," 52-53
 Cavendish, H., 321, 323
 Central tendency, measures of, 181-189
 Clinical method, evidence report data, 57-58
 Cochran, W. G., 211*n*, 355, 363
 Cohen, M. R., 50
 Cole, R. E., 110*n*
 Concatenation, 324
 "Confidence interval estimation," term, 329*n*
 Confounding, 121-122, 136, 150-152
 Control conditions:
 abandonment of experiment, 125
 asymmetrical transfer, 134-135
 balancing, 128-132
 constancy of conditions, 126-128
 control problems, 141-143
 counterbalancing, 132-135
 differential transfer, 134-135
 elimination, 126
 experimenter influences, 139-141
 extraneous variables:
 confounding, 121-122, 136
 control, 137-139
 controllable, 124
 defined, 121
 determination of, 123-124
 experimenter influences, 139-141
 uncontrollable, 124-125
 independent variables, 120-121, 149-152
 nature of, 119-123
 randomization, 135-137
 techniques, 125-137
 Cook, B. S., 17
 Copernicus, 320, 322, 325
 Cornsweet, T. N., 67*n*
 Correlation, meaning of, 166-168
 Correlation coefficient, computation of, 168-170
 Counterbalanced Design for Three Independent Variables (table); 134
 Counterbalancing to Control an Extraneous Variable (table), 133
 Covert Oral Behavior During Various Activities (table), 220
 Cox, G. M., 211*n*, 355, 363
 Crandell, S. A., 218
 Crucial experiment, term, 62
 Cumulative response curve, 299-302
- Dalal, S. N., 362
 Dallenbach, K. M., 29-30
 Data:
 analysis, accuracy of, 349-352
 obtaining of, 28-31 (*see also* Problems)
 retainment of original, 352-353
 statistical treatment, 75
 summary, in experiment results write-up, 85-89
 Deese, J., 29
 Definitions, inadequate, 25-28 (*see also* Problems)
 Dependent variables, 152-161
 Design, in experiment write-up, 84-85
 Design, factorial (*see* Factorial design)
 Design, multi-group (*see* Multi-group design)
 Design selection:
 borderline significance, 115
 empirical hypothesis, 110
 "equality" of groups through randomization, 97-99
 in experimental plan, 70-71
 factorial design (*see* Factorial design)
 general orientation, 96-117
 level of significance, 109*n*
 more than two groups (*see* Multi-group design)
 "null hypotheses," 105-106, 107, 109-110, 113-115
 randomized groups vs. matched groups, 178-180
 significant figures, 102*n*
 statistical analysis, 99-115
 systematic observation method, 115-116
 t test, 100-117
 table of *t*, 108
 two-matched-group (*see* Two-matched-group design)
 "two-tailed test," 105
 Dictionary habit, 93
 Dixon, W. J., 355
 Dodge, R., 44
 "Dollennayer Happiness Scale," 153
 Doré, L. R., 17
 Doyle, A. C., 318-319, 322-323
 Drive (table), 174
 Drive Level (table), 175

- Dubs, H. H., 20, 48
Duncan, D. B., 206*n*, 218
Duncan's Range Test, 204-210, 218, 240-243, 357
- Eason, N., 336
Ebbinghaus, H., 44, 290, 295-296
Edwards, A. L., 134, 146*n*, 160*n*, 189, 194, 206*n*, 356
Edwards, W., 257, 315*n*
Einstein, A., 48, 321, 322, 323, 324, 325
Empirical hypothesis, 37, 110
Engram, W. C., 335
Error variance, reduction of, 356-363
Evidence reports:
 in experimental plan, 56-61
 formulation, 75-76, 304-310
 and hypotheses, 76, 303-317
 irrelevant, 313
 term, 304*n*
Existential hypotheses, 54, 311-312
Expected Mean Squares for the Fixed, Random, and Mixed Models for a 2×2 Factorial Design (table), 286
Experiment, defined, 59*n*
"Experimental Analysis of Behavior," 299
Experimental control (*see* Control conditions)
Experimental design (*see* Design selection)
Experimental methods, 58-61
Experimental plan (*see* Plan of experiments)
Experimental procedure, 74-75
Experimentation, general view, 1-14
Experiments, writing up, 82-94
Explanation, concept of, 324-325
Explanation, in experimentation, 344-347
- F*-test, 229-238, 280-283, 350*n*, 362
Factorial design:
 analysis of variance, 259-271, 280-283
 Diagram, 246
 error terms, 271-273, 285-288
 in experimental plan, 70-71
 F-test for a 2×2 design, 280-283
 fixed models, 271-272
 generalizations, 332-333
 illustration of, 245-249
 interactions, 273-277
 interpreting the *F*'s, 270-271
 the $K \times L$ design, 255-256
 the $K \times L \times M$ design, 258
 mixed models, 273
 random models, 272-273
 statistical analysis, 259-271
 three or more independent variables, 256-258
- Factorial design (*cont.*)
 two independent variables, 253-256
 the 2×2 design, 253-254
 the $2 \times 2 \times 2$ design, 256-258
 the 3×2 design, 254
 the 3×3 design, 255
 the $6 \times 5 \times 4 \times 2$ design (table), 332
 types of, 253-258
 value of, 277-280
- Falling Bodies, Law of, 320, 322, 323, 324
Farber, I. E., 173, 348
Feasting Philosophers, 120
Feigl, H., 21*n*, 347
Field study method, evidence report data, 58
Figures, construction, 87-89
Figures, definition, 86
Fisher, A. E., 52-53
Fisher, R. A., 108, 114*n*, 229, 237, 278-279, 363
Five-groups design, 210-218
Frank, P. G., 21*n*, 28*n*
Freeman, P. R., 135
Freud, S., 26
Fried, C., 226
- G Ratios for Three Methods of Studying a Programmed Text (table), 205
Gaito, J., 135, 240*n*, 298
Galileo, 320, 322, 323, 324, 325
Gampel, D. H., 256
Generalizations:
 "confidence interval estimation," 329*n*
 in experimentation, 327-344
 factorial designs, 332-333
 of findings, 76-77
 limitations, 333-344
 mechanics of, 328-333
 scientific, 322
- Grant, D. A., 89*n*, 94, 315
Gravitation, law of, 321, 322, 323, 324, 325, 326
Grice, G. R., 296-297
Grings, W. W., 67*n*
Guilford, J. P., 89, 182*n*
Guthrie, E. R., 31
- Hall, J. B., 110*n*
Hammond, K. R., 329*n*
Handwriting, 292
Harley, W. F., Jr., 245, 265, 266
Harley, W. F., Sr., 245, 265, 266
Harlow, H. F., 111, 113
Harris, F. R., 300-301, 302
Hays, W. L., 206*n*
Heliocentric Theory of Planetary Motion, 320, 322, 325

- Hempel, C. G., 21n, 41, 44n, 304n, 344
 Hernández-Péon, R., 31
 Herodotus, 120
 Hess, E. H., 66
 Hilgard, E. R., 17
 Hofman, A., 154
 Holland, C. H., 211
 Holmes and Watson, 318-319, 322-323
 Horowitz, S. D., 226
 Horsnell, G., 356
 Hull, C. L., 20, 43, 44,*58
 Hunter, J. J., 296-297
 Hutchens, C., 336
 Hypotheses:
 accidental occurrences, 52-55
 analytic statements, 37-39
 arriving at, 46-49
 characteristics of, 20
 contradictory statements, 37-39
 criteria of value, 49-50
 defined, 37
 empirical, 37, 110
 evidence reports, 303-317
 existential, 54, 311-312
 "facts," 36-37
 guidance function, 51
 "hypothesis myopia," 53-54
 irrelevant evidence reports, 313
 limited, 313
 manner of stating, 40-43
 nature of, 35-37
 null, 105-106, 107, 109-110, 113-115
 parsimony, principle of, 49-50
 Possible Kinds of Statements (table), 38
 research, 53-54
 serendipity, 52-55
 statement of, 40-43, 65
 synthetic statements, 37-39
 truth values, 37-38
 types of, 43-46
 universal, 310-311
 Hysteria Scale of Minnesota Multiphasic
 Personality Inventory, 340
 Inductive schema:
 concatenation, 324
 explanation, concept of, 324-325
 generalizations, 322
 inferences, 322-323
 prediction, 325-326
 Igel, G. J., 110, 111
 Independent variables, 147-152
 Inferences:
 confirmation vs. verification, 309-310
 direct vs. indirect statements, 307-309
 evidence reports, formation of, 304-310
 existential hypotheses, 311-312
 Inferences (*cont.*)
 hypotheses and evidence reports, 303-317
 "if-then" form, 313
 inductive and deductive, 305-307
 and inductive schema, 322-323
 irrelevant evidence reports, 313
 limited hypotheses, 313
 probability inference, term, 306
 unilateral verifiability, 311
 universal hypotheses, 310-311
 Inferences, inductive and deductive, 322-323 (*see also* Inductive schema)
 Interaction, concept of, 249-253, 273-277
 Diagram, 252
 table, 253
 Introduction, in experiment write-up, 83-84
 Jacobson, A. L., 101, 226, 231, 295
 James, W., 111
 Jenkins, J. G., 29-30
 Johnson, P., 258
 Johnson, R. W., 352
 Jones, F. P., 120
 "The Journal of Crazy Ideas" (proposed), 33
Journal of Experimental Psychology, 89n, 92, 139
 Jouvett, M., 31
 Jung, C. G., 26
 Kepler, J., 320, 322, 323, 325
 Kiesler, D. J., 14n, 278
 Kramer, C. Y., 218
 Labeling experiments, 64
 Lachman, R., 266-267
 "Law of the Hammer," 69-70
 Learning, Guthrie's principle of, 31
 Leibniz, G. W., 50
 Lepley, W. M., 291
 Leverrier, U. J. J., 321
 Li, J. C. R., 206n
 Life-history method, evidence report data, 57-58
 Limited hypotheses, 313
 Lindquist, E. F., 72n, 160, 273n, 353, 355
 Literature survey, 64
 Locke, John, 48
 Look vs. Bet (table), 257
 Factorial Design as a Single Two-Groups
 Experiment (table), 249
 Lower, J. S., 315n
 MacCaslin, E. F., 98, 128, 180
 McConnell, J. V., 339

- McFann, H. H., 348
- McGuigan, F. J., 45, 98, 128, 139, 172, 180, 205*n*, 218, 291, 314, 336, 337, 343
- McKeachie, W. J., 41
- McNemar, Q., 89, 189, 355
- MAS (*see* Taylor Manifest Anxiety Scale)
- Magsdick, W. K., 29
- Mann-Whitney U test, 110*n*-
- Martin, I., 67*n*
- Massey, F. J., Jr., 355
- Matched groups vs. randomized groups design, 178-180
- Matching variable, selection of, two-matched-group design, 170-172
- Median Trials to Criterion on Successive Visual Pattern Discriminations (table), 87
- Meehan, J. T., 266-267
- Meeker, M. R., 342
- Mendel, G., 33-34
- Mercury, perihelion of, 321, 323, 326
- Method, in experiment write-up, 84-85
- "The Methodological Character of Theoretical Concepts," 21*n*
- Miller, H. L., 315*n*
- Minnesota Multiphasic Personality Inventory, Hysteria Scale of, 340
- Morgan, Lloyd, 50
- Mosteller, F., 353
- Mowrer, O. H., 171
- Multi-group design:
- analysis of variance, 222-238
 - Duncan's Range Test, 204-210, 218, 240-243, 357
 - in experimental plan, 70
 - F-test, 229-238, 280-283, 350*n*, 362
 - five groups design, 210-218
 - G ratio, 205
 - inductive simplicity, 199
 - mean squares, 228-229
 - statistical analysis of, 203-238
 - statistical analysis for unequal *n*'s, 218-222
 - total sum of squares value, 223-224
 - value of using, 193-203
- Nagel, E., 50
- Nägeli, Carl von, 33-34
- Negative results, experiment write-up, 90
- Neptunus, discovery of, 321, 326
- Newbury, E., 50
- Newton, I., 48, 321, 322, 323, 324, 325
- Nonexperimental methods, evidence report data, 57-58
- Null hypotheses, 105-106, 107, 109-110, 113-115
- Null hypotheses, rejection of, 356-363
- Number of subjects, per group, 363-365
- Occam's razor, 50
- Operationism, 28*n*
- Oppenheim, P., 41, 344
- Organismic variables, 148-149
- Original data, retainment of, 352-353
- Ostrov, N. H., 291
- Overall, J. E., 362
- Page, I. H., 154
- Parametric tests, 110*n*
- Pavlov, I. P., 154, 155
- Pearson Product Moment Coefficient of Correlation, 166, 168
- Peters, C. C., 171*n*
- Peters, R. J., Jr., 205*n*
- Pilot experiments, 62-63, 75
- Plan of experiments:
- abstract section, in write-up, 91
 - authorship, in write-up, 83
 - conducting an experiment, 78-94
 - designing experiments, 63-64
 - design write-up, 84-85
 - discussion write-up, 90
 - "do's" and "don'ts" in write-up, 91-94
 - evidence reports, 56-61
 - experimental methods, 58-61
 - introduction, in write-up, 83-84
 - method write-up, 84-85
 - negative results, write-up, 90
 - nonexperimental methods, 57-58
 - outline for, 64-77
 - procedure write-up, 85
 - references, write-up, 90-91
 - results, write-up, 85-89
 - statistical approach, 71-75
 - summary check list, 77-78
 - title, in write-up, 83
 - types of experiments, 61-63
 - write-up "do's" and "don'ts," 91-94
 - writing up experiments, 82-94
- Planarian (table), 227
- Planetary Motion, Heliocentric Theory of, 320, 322, 325
- Planetary Orbits, laws of, 320, 322
- Pollie, D., 41
- Population, term, 71
- Population study, 71-74
- Poulton, E. C., 135
- Precision, of design, 361
- Prediction, in experimentation, 347-348
- Prediction, in science, 325-326
- Problems:
- absence of information, 16-17
 - contradictory results, 17-19

- Problems (*cont.*)
 data, obtaining of, 28-31
 explaining facts, 19-20
 forgetting, theories of, 29-30
 gaps in knowledge, 16-17
 Guthrie's principle of learning, 31
 hypotheses, 20
 impasse problems, 32
 importance criterion, 31-33
 inadequate definitions, 25-28
 manifestations, 15-16
 nature of, 15
 possibilities, kinds of, 23
 pseudoquestions, 21*n*
 psychological reactions to, 33-34
 reminiscence, 29
 solvable, 21
 statement of, 65
 testability:
 application of criteria, 24-25
 classes of, 23-24
 probability theory, 22
 truth theory, 21-22
 unstructured, 25
 vicious circularity, in reasoning, 30-31
 working principle, 24
- Psammetichos, king of Egypt, 120
- Pseudoquestions, 21*n* (*see also* Problems)
- Psychasthenic Scale, 340
- Psychological Abstracts*, 64
- Psychological experiment, example, 12-14
- Psychotherapy Research and the Search for a Paradigm, 14*n*
- Ptolemaic theory, 320
- Publication Manual*, American Psychological Association, 83
- Quantum mechanics, 321
- Randomization technique, 72-74, 135-137
- Randomized groups, 95-118 (*see also* Design selection)
- Randomized groups vs. matched groups design, 178-180
- Ray, W. S., 206*n*, 361
- References, experiment write-up, 90-91
- Reichenbach, H., 21*n*, 40, 44*n*, 46, 199, 311, 314, 319, 320, 321
- Relativity, theory of, 321, 322, 323, 324, 325
- Reminiscence, 29
- Renshaw, S., 17
- Research, 53 (*see also* Hypotheses)
- Response variables, 148
- Results write-up, 85-89
- Reynolds, R. W., 342
- Reynolds, T., 336
- Richardson, E. C., 172
- Rosenthal, 140
- Rosenthal, R., 140-141
- Rotberg, I. C., 254
- Russell, B., 40-41
- Ryan, T. A., 240
- Sandler, J., 177
- Saucer, R. T., 67
- Savukas, R. A., 291
- Scherrer, H., 31
- Schlosberg, H., 58-59, 146*n*, 290
- Schoenfeld, W. N., 201
- Schrödinger, E., 321
- Schwarzbeck, W. C., 17
- Science, nature of, 1-4
- Scientific method, in psychological experiments, 4-12
- Scientific progress, general form of, 318-326
- Scriven, M., 11*n*, 21*n*, 344*n*
- Seidel, R. J., 254
- "Serendip, Three Princes of," 52
- Sidman, M., 53
- Sidowski, J. B., 67*n*
- Siegel, S., 356
- Silverman, H. F., 41
- Singh, S. D., 86-88
- Sjoholm, N. A., 339
- Skinner, B. F., 60, 154*n*, 299-301
- Skinner Box, 47-48, 126, 153, 201, 299, 301, 343
- Snedecor, G. W., 229
- Socrates, 345
- Solomon, R. L., 131
- Spearman Rank Correlation Coefficient, 169-170
- Speisman, J., 41
- Spence, K. W., 146, 149*n*, 173, 348
- Spires, A. M., 340-341
- Standard deviation, 181-189
- Statistical analysis:
 A-test, 176-178
 assumptions, 354-356
 factorial designs, 259-271
 multi-group design, 203-238
 tests, 75, 354-356
 treatment of data, 75
- Statistical Methods for Research Workers*, 108
- Statistical Tables of Biological, Agricultural, and Medical Research*, 237
- Steadiness (table), 131
- Steadiness Scores of Soldiers Before and After Rifle Training (table), 128
- Steadiness Scores of Trained and Untrained Groups (table), 129

- Stimulus variables, 148
- Subjects:
- experiment write-up information, 84
 - number per group, 363-365
 - selection and assignment, 71-74
- Suiter, R. D., 218
- Sulzbacher, S. I., 204-205
- Sums of Squares and df for the 2×2 Factorial Design, 263
- Sums of Squares for the 2×2 Factorial Design (table), 262
- Systematic observation, evidence report data, 58
- t test, 100-117, 165-166, 171, 176-177, 180-181, 223, 224, 229, 230, 231, 238-240, 356-358, 364-365
- Table of A , 177
- Table of F , 232-237
- Table of t , 108
- Tables, construction, 86-87, 89
- Tables, defined, 86
- Tape recorders, use of, 359-360
- Taylor, J. A., 114, 173, 345
- Taylor Manifest Anxiety Scale (MAS), 114, 173, 345
- Temperature of water (table), 211
- Testability (*see also* Problems):
- application of criterions, 24-25
 - classes of, 23-24
 - probability theory, 22
 - truth theory, 21-22
- Tests of significance, combining of, 353-354
- Thorndike, E. L., 5, 25
- 3×2 Factorial Design (table), 254
- Tides-Moon Law, 320-321, 322, 325
- Title, in experiment write-up, 83
- Tversky, A., 257
- "Two-groups" design, 95-118 (*see also* Design selection)
- Two-Groups Design in which the Data are Analyzed as a Function of Two Experimenters (table), 335
- Two-matched-group design:
- computation of t , 189-190
 - correlation, 166-172
 - examples of, 163-164, 172-176
 - r_{12} symbol, 184-188
 - randomized groups vs. matched groups, 178-180
 - selection of matching variable, 170-172
 - standard deviation and variance, 181-189
 - statistical analysis of, 164-166
 - statistical analysis with A -test, 176-178
 - variance of a set of values, 183
- 2×2 Factorial Design with Fictitious Means (table), 334
- 2×2 Factorial Design with Strength of Word Association and Observing Response as the Two Variables (table), 268
- Underwood, B. J., 28 n , 48, 51 n , 62, 134, 135, 141, 146 n , 222 n , 290, 293, 295-296
- "Unified field" theory, 321
- Universal hypotheses, 310-311
- Values of r_p for Duncan's Range Test (table), 207, 217
- Values of r_p for a Five-Groups Design with 31 df (table), 221
- Values of r_p and R_p for Five Groups with 45 df (table), 213
- Values of r_p and R_p for 2 and 3 Groups with 57 df (table), 208
- Van Voorhis, W. R., 171 n
- Variability, measures of, 181-189
- Variables:
- confounding, 150-152
 - control, 70
 - control of independent variable, 149-152
 - definition, 65
 - dependent variables:
 - delayed measurements, 161
 - growth measures, 160
 - measuring of, 152-153
 - multiple, 159-160
 - reliability, 157-159
 - selection, 154-159
 - time changes, 160, 161
 - trend analysis, 160 n
 - validity, 155-157
 - independent variables, 147-152
 - organismic variables, 148-149
 - quasi-experiments, 152
 - relationships, types of, 144-147
 - response variables, 148
 - stimulus variables, 148
- Variance, 181-189
- Venables, P. H., 67 n
- Vicious circularity, in reasoning, 30-31 (*see also* Problems)
- Walpole, H., 52
- War of the Worlds*, radio dramatization, 58
- Weber's Law, 199 n , 290
- Webster, R. L., 18
- Weingold, H. P., 18
- Welles, O., 58

- Wells, H. G., 58
 Wertheimer, M., 19*n*
 Whipple Steadiness Test, 27
 William of Occam, 50
 Wilson, W., 315*n*
 Wine, R. L., 273
 Winer, B. J., 206*n*, 273*n*
 Within-subjects designs:
 cumulative response curve, 299-302
 evaluation, 294-298
 F-test, 298-299
 several conditions, many subjects, 293-294
 Within-subjects designs (*cont.*)
 single-subject with replication, 298-301
 two conditions, many subjects, 290-292
 Wolf, M. M., 300-301
 Wolins, L., 350*n*, 353
 Woods, P. J., 211, 335
 Woodward, W. T., 110*n*
 Woodworth, R. S., 58-59, 146*n*, 290
 Writing up experiments, 82-94
 Yanof, H. M., 67*n*
 Yates, 237
 Zimmerman, R. R., 111